nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 1.2 [Juni 2008]

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 1.2
Juni 2008

Herausgeber

Heike Neuroth
Hans Liegmann †
Achim Oßwald
Regine Scheffel
Mathias Jehn
Stefan Strathmann

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Im Auftrag von

nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de

Kontakt

editors@langzeitarchivierung.de
c/o
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Dr. Heike Neuroth
Forschung und Entwicklung
Papendiek 14
37073 Göttingen
Tel. +49 (0) 55 1 39 38 66

Der Inhalt steht unter folgender Creative Commons Lizenz:
http://creativecommons.org/licenses/by-nc-sa/2.0/de/

Vorwort **9**

1 Einführung **1-1**

2 Rechtliche Aspekte **2-1**

3 State of the Art **3-1**

LZA-Aktivitäten in Deutschland aus dem Blickwinkel von nestor 3-1
3.1 Bibliotheken.....3-8
3.2 Archive.....3-11
3.3 Museen.....3-14

4 Rahmenbedingungen für die Langzeitarchivierung digitaler Objekte **4-1**

4.1 Nationale Preservation Policy.....4-3
4.2 Institutionelle Preservation Policy.....4-6
4.4 Auswahlkriterien.....4-10

5 Geschäftsmodelle **5-1**

5.1 Kosten.....5-1
5.2 Service- und Lizenzmodelle.....5-7

6 Organisation **6-1**

7 Das Referenzmodell OAIS - Open Archival Information System **7-1**

8 Vertrauenswürdigkeit von digitalen Langzeitarchiven **8-1**

8.1 Grundkonzepte der Sicherheit und Vertrauenswürdigkeit digitaler Objekte

.....8-2
8.2 Praktische Sicherheitskonzepte.....8-5
8.3 Evaluierung der Vertrauenswürdigkeit digitaler Archive.....8-15
8.4 Literatur.....8-24

9 Formate 9-1

Einleitung.....9-1
9.1 Digitale Objekte.....9-3
9.2 Dateiformate.....9-7
9.4 Formaterkennung und Validierung.....9-9
9.5 File Format Registries.....9-11
9.6 Tools.....9-15

10 Standards und Standardisierungsbemühungen 10-1

10.1.1 Metadata Encoding and Transmission Standard: Das METS Abstract Model – Einführung und Nutzungsmöglichkeiten.....10-1
10.1.3 PREMIS.....10-7
10.1.4 LMER.....10-11
10.1.5 MIX.....10-14

11 Hardware 11-1

11.1 Hardware-Environment.....11-1
11.2 Digitale Speichermedien.....11-4
11.2.1 Magnetbänder.....11-8
11.2.2 Festplatten.....11-14

12 Digitale Erhaltungsstrategien 12-1

Einleitung.....12-1
12.1 Bitstream Preservation.....12-3
12.2 Migration.....12-10
12.3 Emulation.....12-16
12.4 Computermuseum.....12-24

12.5 Mikroverfilmung 12-31

13 Access 13-1

13.1 Retrieval 13-3

13.2 Persistent Identifier (PI) - ein Überblick 13-6

13.2.1 Der Uniform Resource Name (URN) 13-23

13.2.2 Der Digital Objekt Identifier (DOI) und die Verwendung zum Primärdaten-Management 13-36

14 Technischer Workflow 14-1

14.1 Einführende Bemerkungen und Begriffsklärungen 14-1

14.2 Workflow in der Langzeitarchivierung: Methode und Herangehensweise 14-4

14.3 Technisches Workflowmanagement in der Praxis: Erfahrungen und Ergebnisse 14-8

15 Anwendungsfelder in der Praxis 15-1

Einleitung 15-1

15.1 Textdokumente 15-3

15.2 Bilddokumente 15-8

15.3 Multimedia/Komplexe Applikationen 15-15

15.3.2 Audio 15-20

15.3.3 Langzeitarchivierung und -bereitstellung im E-Learning-Kontext 15-23

15.3.4 Interaktive Applikationen 15-28

15.4 Web-Harvesting zur Langzeiterhaltung von Internet-Dokumenten 15-42

15.5 Wissenschaftliche Primärdaten 15-52

15.6 Computerspiele 15-61

16 Lernen und weitergeben – Aus- und Weiterbildungsangebote zur Langzeitarchivierung 16-1

Vorwort

Liebe Leserinnen und Leser,

wir freuen uns, Ihnen die zweite, aktualisierte online Ausgabe (Version 1.2, Juni 2008) des nestor Handbuchs „Eine kleine Enzyklopädie der digitalen Langzeitarchivierung“ präsentieren zu können.
Das nestor Handbuch will nach dem Konzept des „Living Document“ das derzeitige Wissen über das vielfältige und komplexe Thema der Langzeitarchivierung und Langzeitverfügbarkeit digitaler Objekte und seine unterschiedlichen Teilaspekte sammeln und über eine „kleine Enzyklopädie“ in strukturierter Form den Interessierten in deutscher Sprache zugänglich machen.
Einzelne, von verschiedenen Experten erstellte Fachbeiträge gestatten einen Überblick, manchmal auch einen vertieften Einblick in die diversen Themengebiete der Langzeitarchivierung: von technischen und rechtlichen Aspekten bis hin zur Definition von Rahmenbedingungen.
In dieser Version des Handbuchs finden Sie neben neu hinzu gekommenen Kapiteln bzw. aktualisierten Beiträgen nun auch die Gelegenheit, mit dem Autor/der Autorin direkt Kontakt aufzunehmen und Ihre Kommentare und Ergänzungen einzubringen.
Wir hoffen, dass Sie davon regen Gebrauch machen. Nur durch Ihre aktive Mitarbeit wird das Konzept eines „Living Document“ verwirklicht und fließen

aktuelle Entwicklungen sowie unterschiedliche Sichten rasch in das Handbuch ein. Die kleine Enzyklopädie soll somit zu einem unentbehrlichen Gebrauchsinstrument sowohl für Fachkolleginnen und Fachkollegen im Rahmen ihrer alltäglichen beruflichen Praxis als auch für interessierte Laien werden.

Wir freuen uns, dass wir eine große Anzahl an Autoren gewinnen konnten, für deren Beiträge wir uns auf diesem Weg ganz herzlich bedanken.
Ein großer Dank gilt auch den anderen Miteditoren für die redaktionelle Betreuung und inhaltliche Koordinierung der Artikel.

Gestatten Sie mir an dieser Stelle – auch im Namen der anderen Editoren – unseres langjährigen Kollegen und bisherigen Mitherausgeber Hans Liegmann zu gedenken, der im November 2007 im Alter von nur 54 Jahren tödlich verunglückt ist. Er war einer der Initiatoren des Projektes nestor und hat mit hohem Engagement, ausgewiesener Fachkompetenz und charmanter Kollegialität die Aktivitäten des Editorial Board begleitet. Wir vermissen ihn schmerzlich!
Das Andenken an ihn motiviert, das von ihm mit Begonnene auch in seinem Sinne weiter zu entwickeln.

Allen Lesern wünsche ich viel Freude bei der Lektüre des Handbuchs.
Ergänzungen, Anmerkungen und Korrekturen sind auch weiterhin willkommen!

Beste Grüße,
Ihre Heike Neuroth

1 Einführung

Hans Liegmann, Heike Neuroth

1. Die digitale Welt, eine ständig wachsende Herausforderung

Die Überlieferung des kulturellen Erbes, traditionell eine der Aufgaben von Bibliotheken, Archiven und Museen, ist durch die Informationstechnologien deutlich schwieriger geworden.

In der heutigen Zeit werden zunehmend mehr Informationen digital erstellt und veröffentlicht. Diese digitalen Informationen, die Güter des Informations- und Wissenszeitalterzeitalter, sind einerseits wertvolle kulturelle und wissenschaftliche Ressourcen, andererseits sind sie sehr vergänglich. Die Datenträger sind ebenso der Alterung unterworfen, wie die Datenformate oder die zur Darstellung notwendige Hard- und Software. Um langfristig die Nutzbarkeit der digitalen Güter sicherzustellen, muss schon frühzeitig Vorsorge getroffen werden, müssen Strategien der digitalen Langzeitarchivierung entwickelt und umgesetzt werden.

Die Menge und die Heterogenität der Informationen, die originär in digitaler Form vorliegen, wächst beständig an.

In großem Umfang werden ursprünglich analog vorliegende Daten digitalisiert (z.B. Google Print Projekt[1]), um den Benutzerzugriff über Datennetze zu ver-

1 http://print.google.com

einfachen. Im Tagesgeschäft von Behörden, Institutionen und Unternehmen werden digitale Akten produziert, für die kein analoges Äquivalent mehr zur Verfügung steht.
Sowohl die wissenschaftliche Fachkommunikation wie der alltägliche Informationsaustausch sind ohne die Vermittlung von Daten in digitaler Form nicht mehr vorstellbar.
Mit der Menge der ausschließlich digital vorliegenden Information wächst unmittelbar auch ihre Relevanz als Bestandteil unserer kulturellen und wissenschaftlichen Überlieferung sowie die Bedeutung ihrer dauerhaften Verfügbarkeit für Wissenschaft und Forschung. Denn das in der „scientific community" erarbeitete Wissen muss, soll es der Forschung dienen, langfristig verfügbar gehalten werden, da der Wissenschaftsprozess immer wieder eine Neubewertung langfristig archivierter Fakten erforderlich macht. Die Langzeitarchivierung digitaler Ressourcen ist daher eine wesentliche Bedingung für die Konkurrenzfähigkeit des Bildungs- und Wissenschaftssystems und der Wirtschaft. In Deutschland existiert eine Reihe von Institutionen (Archive, Bibliotheken, Museen), die sich in einer dezentralen und arbeitsteiligen Struktur dieser Aufgabe widmen.
Im Hinblick auf die heutige Situation, in der Autoren und wissenschaftliche Institutionen (Universitäten, Forschungsinstitute, Akademien) mehr und mehr selbst die Veröffentlichung und Verbreitung von digitalen Publikationen übernehmen, erscheint auch weiterhin ein verteilter Ansatz angemessen, der jedoch um neue Verantwortliche, die an der „neuen" Publikationskette beteiligt sind, erweitert werden muss.

1.1. Langzeitarchivierung im digitalen Kontext

„Langzeitarchivierung" meint in diesem Zusammenhang mehr als die Erfüllung gesetzlicher Vorgaben über Zeitspannen, während der steuerlich relevante tabellarisch strukturierte Daten verfügbar gehalten werden müssen. „Langzeit" ist die Umschreibung eines nicht näher fixierten Zeitraumes, währenddessen wesentliche, nicht vorhersehbare technologische und soziokulturelle Veränderungen eintreten; Veränderungen, die sowohl die Gestalt als auch die Nutzungssituation digitaler Ressourcen in rasanten Entwicklungszyklen vollständig umwälzen können. Es gilt also, jeweils geeignete Strategien für bestimmte digitale Sammlungen zu entwickeln, die je nach Bedarf und zukünftigem Nutzungsszenarium die langfristige Verfügbarkeit der digitalen Objekte sicherstellen. Dabei spielen nach bisheriger Erfahrung das Nutzerinteresse der Auf- und Abwärtskompatibilität alter und neuer Systemumgebungen nur dann eine Rolle, wenn dies dem Anbieter für die Positionierung am Markt erforderlich erscheint.

„Langzeit“ bedeutet für die Bestandserhaltung digitaler Ressourcen nicht die Abgabe einer Garantieerklärung über fünf oder fünfzig Jahre, sondern die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können.
Der Bedeutungsinhalt von „Archivierung“ müsste hier nicht näher präzisiert werden, wäre er nicht im allgemeinen Sprachgebrauch mit der fortschreitenden Anwendung der Informationstechnik seines Sinnes nahezu entleert worden. „Archivieren“ bedeutet zumindest für Archive, Museen und Bibliotheken mehr als nur die dauerhafte Speicherung digitaler Informationen auf einem Datenträger. Vielmehr schließt es die Erhaltung der dauerhaften Verfügbarkeit digitaler Ressourcen mit ein.

2. Substanzerhaltung

Eines von zwei Teilzielen eines Bestandserhaltungskonzeptes für digitale Ressourcen ist die unversehrte und unverfälschte Bewahrung des digitalen Datenstroms: die Substanzerhaltung der Dateninhalte, aus denen digitale Objekte physikalisch bestehen. Erfolgreich ist dieses Teilziel dann, wenn die aus heterogenen Quellen stammenden und auf unterschiedlichsten Trägern vorliegenden Objekte möglichst früh von ihren originalen Träger getrennt und in ein homogenes Speichersystem überführt werden. Die verantwortliche archivierende Institution wird vorzugsweise ein funktional autonomes Teilsystem einrichten, dessen vorrangige Aufgabe die Substanzerhaltung digitaler Ressourcen ist. Wichtige Bestandteile dieses Systems sind automatisierte Kontrollmechanismen, die den kontinuierlichen systeminternen Datentransfer überwachen. Die kurze Halbwertszeit technischer Plattformen macht auch vor diesem System nicht halt und zwingt zum laufenden Wechsel von Datenträgergenerationen und der damit möglicherweise verbundenen Migration der Datenbestände.
Dauerhafte Substanzerhaltung ist nicht möglich, wenn die Datensubstanz untrennbar an einen Datenträger und damit an dessen Schicksal gebunden ist. Technische Maßnahmen zum Schutz der Verwertungsrechte (z.B. Kopierschutzverfahren) führen typischerweise mittelfristig solche Konfliktsituationen herbei. Ein digitales Archiv wird in Zukunft im eigenen Interesse Verantwortung nur für solche digitalen Ressourcen übernehmen, deren Datensubstanz es voraussichtlich erhalten kann. Ein objektspezifischer „Archivierungsstatus“ ist in dieser Situation zur Herstellung von Transparenz hilfreich.

3. Erhaltung der Benutzbarkeit

Substanzerhaltung ist nur eine der Voraussetzungen, um die Verfügbarkeit und Benutzbarkeit digitaler Ressourcen in Zukunft zu gewährleisten. „Erhaltung der Benutzbarkeit" digitaler Ressourcen ist eine um ein Vielfaches komplexere Aufgabenstellung als die Erhaltung der Datensubstanz. Folgen wir dem Szenario eines „Depotsystems für digitale Objekte", in dem Datenströme sicher gespeichert und über die Veränderungen der technischen Umgebung hinweg aufbewahrt werden, so steht der Benutzer/die Benutzerin der Zukunft gleichwohl vor einem Problem. Er oder sie ist ohne weitere Unterstützung nicht in der Lage den archivierten Datenstrom zu interpretieren, da die erforderlichen technischen Nutzungsumgebungen (Betriebssysteme, Anwendungsprogramme) längst nicht mehr verfügbar sind. Zur Lösung dieses Problems werden unterschiedliche Strategien diskutiert, prototypisch implementiert und erprobt.
Der Ansatz, Systemumgebungen in Hard- und Software-Museen zu konservieren und ständig verfügbar zu halten, wird nicht ernsthaft verfolgt. Dagegen ist die Anwendung von Migrationsverfahren bereits für die Substanzerhaltung digitaler Daten erprobt, wenn es um einfachere Datenstrukturen oder den Generationswechsel von Datenträgertypen geht. Komplexe digitale Objekte entziehen sich jedoch der Migrationsstrategie, da der für viele Einzelfälle zu erbringende Aufwand unkalkulierbar ist. Aus diesem Grund wird mit Verfahren experimentiert, deren Ziel es ist, Systemumgebungen lauffähig nachzubilden (Emulation). Es werden mehrere Ansätze verfolgt, unter denen die Anwendung formalisierter Beschreibungen von Objektstrukturen und Präsentationsumgebungen eine besondere Rolle einnimmt.
Die bisher genannten Ansätze spielen durchgängig erst zu einem späten Zeitpunkt eine Rolle, zu dem das digitale Objekt mit seinen für die Belange der Langzeitarchivierung günstigen oder weniger günstigen Eigenschaften bereits fertig gestellt ist. Darüber hinaus wirken einige wichtige Initiativen darauf hin, bereits im Entstehungsprozess digitaler Objekte die Verwendung langzeitstabiler Datenformate und offener Standards zu fördern. Welche der genannten Strategien auch angewandt wird, die Erhaltung der Benutzbarkeit und damit der Interpretierbarkeit wird nicht unbedingt mit der Erhaltung der ursprünglichen Ausprägung des „originalen" Objektes korrespondieren. Es wird erforderlich sein, die Bemühungen auf die Kernfunktionen (so genannte „significant properties") digitaler Objekte zu konzentrieren, vordringlich auf das, was ihren wesentlichen Informationsgehalt ausmacht.

4. Technische Metadaten

Die Erhebung und die strukturierte Speicherung technischer Metadaten ist eine wichtige Voraussetzung für die automatisierte Verwaltung und Bearbeitung digitaler Objekte im Interesse ihrer Langzeitarchivierung. Zu den hier relevanten Metadaten gehören z.B. Informationen über die zur Benutzung notwendigen Systemvoraussetzungen hinsichtlich Hardware und Software sowie die eindeutige Bezeichnung und Dokumentation der Datenformate, in denen die Ressource vorliegt. Spätestens zum Zeitpunkt der Archivierung sollte jedes digitale Objekt über einen eindeutigen, beständigen Identifikator (persistent identifier) verfügen, der es unabhängig vom Speicherort über Systemgrenzen und Systemwechsel hinweg identifiziert und dauerhaft nachweisbar macht. Tools, die zurzeit weltweit entwickelt werden, können dabei behilflich sein, Formate beim Ingest-Prozess (Importvorgang in ein Archivsystem) zu validieren und mit notwendigen technischen Metadaten anzureichern. Ein viel versprechender Ansatz ist das JHOVE Werkzeug[2], das zum Beispiel Antworten auf folgende Fragen gibt: Welches Format hat mein digitales Objekt? Das digitale Objekt „behauptet" das Format x zu haben, stimmt dies?
Ohne die Beschreibung eines digitalen Objektes mit technischen Metadaten dürften Strategien zur Langzeitarchivierung wie Migration oder Emulation nahezu unmöglich bzw. deutlich kostenintensiver werden.

5. Vertrauenswürdige digitale Archive

Digitale Archive stehen erst am Beginn der Entwicklung, während Archive für traditionelles Schriftgut über Jahrhunderte hinweg Vertrauen in den Umfang und die Qualität ihrer Aufgabenwahrnehmung schaffen konnten. Es werden deshalb Anstrengungen unternommen, allgemein akzeptierte Leistungskriterien für vertrauenswürdige digitale Archive aufzustellen (vgl. Kap. 8), die bis zur Entwicklung eines Zertifizierungsverfahrens reichen. Die Konformität zum OAIS-Referenzmodell spielt dabei ebenso eine wichtige Rolle, wie die Beständigkeit der institutionellen Struktur, von der das Archiv betrieben wird. Es wird erwartet, dass Arbeitsmethoden und Leistungen der Öffentlichkeit präsentiert werden, sodass aus dem möglichen Vergleich zwischen inhaltlichem Auftrag und tatsächlicher Ausführung eine Vertrauensbasis sowohl aus Nutzersicht, wie auch im Interesse eines arbeitsteiligen kooperativen Systems, entstehen kann.
Wichtig in diesem Zusammenhang ist auch die Wahrung der Integrität und Authentizität eines digitalen Objektes. Nur wenn sichergestellt werden kann, dass

2 JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/

das digitale Objekt zum Beispiel inhaltlich nicht verändert wurde, kann man mit der Ressource vertrauensvoll arbeiten.

6. Verteilte Verantwortung bei der Langzeitarchivierung digitaler Ressourcen

6.1 National

Hinsichtlich der Überlegungen zur Langzeitarchivierung digitaler Quellen in Deutschland muss das Ziel sein, eine Kooperationsstruktur zu entwickeln, die entsprechend den Strukturen im analogen Bereich die Bewahrung und Verfügbarkeit aller digitalen Ressourcen gewährleistet. Diese Strukturen berücksichtigen alle Ressourcen, die in Deutschland, in deutscher Sprache oder über Deutschland erschienen sind, die Bewahrung und Verfügbarkeit der wichtigsten Objekte jedes Fachgebiets organisiert (unabhängig davon, ob es sich um Texte, Fakten, Bilder, Multimedia handelt) sowie die Bewahrung und Verfügbarkeit digitaler Archivalien garantiert.
Das Auffinden der Materialien soll dem interessierten Nutzer ohne besondere Detailkenntnisse möglich sein, d.h. ein weiteres Ziel einer angestrebten Kooperationsstruktur beinhaltet, die Verfügbarkeit durch Zugangsportale zu sicher zu stellen und die Nutzer dorthin zu lenken, wo die Materialien liegen. Dabei müssen selbstverständlich Zugriffsrechte, Kosten u.a. durch entsprechende Mechanismen (z.B. Bezahlsysteme) berücksichtigt werden.
Beim Aufbau einer solchen Struktur sind vor allem die Bibliotheken, Archive und Museen gefordert. In Deutschland müssen in ein entstehendes Kompetenznetzwerk Langzeitarchivierung aber auch die Produzenten digitaler Ressourcen, d. h. Verlage, Universitäten, Forschungseinrichtungen, Wissenschaftler sowie technische Dienstleister wie Rechen-, Daten- und Medienzentren und Großdatenbankbetreiber einbezogen werden.

6.2 Internationale Beispiele

Ein Blick ins Ausland bestärkt den kooperativen Ansatz. In Großbritannien ist im Jahr 2001 die Digital Preservation Coalition (DPC) mit dem Ziel initiiert worden, die Herausforderungen der Langzeitarchivierung und -verfügbarkeit digitaler Quellen aufzugreifen und die Langzeitverfügbarkeit des digitalen Erbes in nationaler und internationaler Zusammenarbeit zu sichern. Die DPC versteht sich als ein Forum, welches Informationen über den gegenwärtigen Forschungsstand sowie Ansätze aus der Praxis digitaler Langzeitarchivierung

dokumentiert und weiterverbreitet. Die Teilnahme an der DPC ist über verschiedene Formen der Mitgliedschaft möglich.
In den USA ist im Jahr 2000 ein Programm zum Aufbau einer nationalen digitalen Informationsinfrastruktur und ein Programm für die Langzeitverfügbarkeit digitaler Ressourcen in der Library of Congress (LoC) verabschiedet worden. Die Aufgaben werden in Kooperation mit Vertretern aus anderen Bibliotheken und der Forschung sowie kommerziellen Einrichtungen gelöst. Darüber hinaus hat die LoC in Folge ihrer Jubiläumskonferenz im Jahre 2000 einen Aktionsplan aufgestellt, um Strategien zum Management von Netzpublikationen durch Bibliothekskataloge und Metadatenanwendungen zu entwickeln. Der Ansatz einer koordinierten nationalen Infrastruktur, auch unter den Rahmenbedingungen einer äußerst leistungsfähigen Nationalbibliothek wie der LoC, bestätigt die allgemeine Einschätzung, dass zentralistische Lösungsansätze den künftigen Aufgaben nicht gerecht werden können.
Im Archivbereich wird die Frage der Langzeitverfügbarkeit digitaler Archivalien in internationalen Projekten angegangen. Besonders zu erwähnen ist das Projekt ERPANET, das ebenfalls den Aufbau eines Kompetenznetzwerks mittels einer Kooperationsplattform zum Ziel hat. InterPares ist ein weiteres internationales Archivprojekt, welches sich mit konkreten Strategien und Verfahren der Langzeitverfügbarkeit digitaler Archivalien befasst. Die Zielsetzung der Projekte aus dem Archivbereich verdeutlichen, wie ähnlich die Herausforderungen der digitalen Welt für alle Informationsanbieter und Bewahrer des kulturellen Erbes sind und lassen Synergieeffekte erwarten.
Ein umfassender Aufgabenbereich von Museen ist das fotografische Dokumentieren und Verfahren von Referenzbildern für Museumsobjekte. Die Sicherung der Langzeitverfügbarkeit der digitalen Bilder ist eine essentielle Aufgabe aller Museen. Im Bereich des Museumswesens muss der Aufbau von Arbeitsstrukturen, die über einzelne Häuser hinausreichen, jedoch erst noch nachhaltig aufgebaut werden.

7. Rechtsfragen

Im Zusammenhang mit der Langzeitarchivierung und -verfügbarkeit digitaler Ressourcen sind urheberrechtlich vor allem folgende Fragestellungen relevant:

- Rechte zur Durchführung notwendiger Eingriffe in die Gestalt der elektronischen Ressourcen im Interesse der Langzeiterhaltung,
- Einschränkungen durch Digital Rights Management Systeme (z. B. Kopierschutz),
- Konditionen des Zugriffs auf die archivierten Ressourcen und deren

Nutzung.

Die EU-Richtlinie zur Harmonisierung des Urheberrechts in Europa greift diese Fragestellungen alle auf; die Umsetzung in nationales Recht muss aber in vielen Ländern, darunter auch Deutschland, noch erfolgen. Erste Schritte sind in dem „ersten Korb" des neuen deutschen Urheberrechtsgesetzes erfolgt.

8. Wissenschaftliche Forschungsdaten

Die Langzeitarchivierung wissenschaftlicher Primär- und Forschungsdaten spielt eine immer größere Rolle. Spätestens seit einigen „Manipulations-Skandalen" (zum Beispiel Süd-Korea im Frühjahr 2008) ist klar geworden, dass auch Forschungsdaten langfristig verfügbar gehalten werden müssen. Verschiedene Stimmen aus wissenschaftlichen Disziplinen, sowohl Geistes- als auch Naturwissenschaften, wünschen sich eine dauerhafte Speicherung und einen langfristigen Zugriff auf ihr wissenschaftliches Kapital.

Weiterhin fordern verschiedene Förderer und andere Institutionen im Sinne „guter wissenschaftlicher Praxis" (DFG) dauerhafte Strategien, wie folgende Beispiele zeigen:

- DFG, Empfehlung 7[3]
- OECD[4]
- Und ganz aktuell die EU[5] mit folgendem Zitat: „Die Europäische Kommission hat am 10. April 2008 die ‚Empfehlungen zum Umgang mit geistigem Eigentum bei Wissenstransfertätigkeiten und für einen Praxiskodex für Hochschulen und andere öffentliche Forschungseinrichtungen' herausgegeben. Zu diesem Thema war bereits im ersten Halbjahr 2007 unter der deutschen Ratspräsidentschaft ein Eckpunktepapier mit dem Titel ‚Initiative zu einer Charta zum Umgang mit geistigem Eigentum an öffentlichen Forschungseinrichtungen und Hochschulen' ausgearbeitet worden."

Es gibt zurzeit in Deutschland konkrete Überlegungen, wie es gelingen kann, gemeinsam mit den Wissenschaftlern eine gute Praxis bezüglich des Umgangs mit Forschungsdaten zu entwickeln. Die beinhaltet auch (aber nicht nur) die Veröffentlichung von Forschungsdaten.

Interessante Fragen in diesem Zusammenhang sind zum Beispiel, wem die Forschungsdaten eigentlich gehören (dem Wissenschaftler, der Hochschule, der

3 http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

4 http://www.oecd.org/dataoecd/9/61/38500813.pdf

5 http://ec.europa.eu/invest-in-research/pdf/ip_recommendation_de.pdf

Öffentlichkeit), was Forschungsdaten eigentlich sind - hier gibt es bestimmt fachspezifische Unterschiede, welche Forschungsdaten langfristig aufbewahrt werden müssen - eine fachliche Selektion kann nur in enger Kooperation mit dem Wissenschaftler erfolgen, und wer für die Beschreibungen z.B. die Lieferung von technischen und deskriptiven Metadaten zuständig ist.

2 Rechtliche Aspekte

Arne Upmeier

Nicht ganz zufällig wird kritisiert, dass die gravierendste Schwäche des Urheberrechts dessen **Kompliziertheit** sei.[1] Das Urheberrecht der digitalen Langzeitarchivierung bildet da keine Ausnahme. Sehr vieles hängt von den konkreten Umständen im Einzelfall ab und lässt sich nicht generalisieren. Die folgenden Ausführungen bleiben daher notwendig allgemein und vieles – im Einzelfall entscheidendes – muss außen vor bleiben.

1. Was darf archiviert werden?

Ein digitales Objekt muss über eine bestimmte Schöpfungshöhe verfügen, um überhaupt im Sinne des Urheberrechts schutzwürdig zu sein, d.h. es muss über einen bestimmten geistigen Inhalt, der in einer bestimmten Form Ausdruck gefunden hat und eine gewisse Individualität verfügen. Nicht jeder Text oder jedes Musikstück unterliegt daher automatisch dem Urheberrecht. Auch eine ungeordnete Sammlung von wissenschaftlichen Rohdaten ist im Regelfall nicht

1 Buck-Heeb, Petra: Stärken und Schwächen des deutschen Urheberrechts in Forschung und Lehre. In: Urheberrecht in digitalisierter Forschung und Lehre. Hrsg. von Nikolaus Forgó, S. 29.

urheberrechtlich geschützt. Digitale Objekte, die danach gar nicht dem Urheberrecht unterliegen, können deswegen im Allgemeinen unproblematisch archiviert werden.

Rechtlich unproblematisch sind auch Dokumente, die aus dem einen oder anderen Grunde **gemeinfrei** sind. Hierzu zählen beispielsweise amtliche Werke § 5 Urheberrechtsgesetz (UrhG), wie etwa Gesetze oder Verordnungen und auch alle Werke, deren Urheberrechtsschutz bereits abgelaufen ist. Dies ist in der Regel siebzig Jahre nach dem Tode des Urhebers der Fall (§ 64 UrhG).[2]

Gesetzlich bisher nicht geregelt ist der Umgang mit sogenannten „verwaisten Werken" (*orphan works*) bei denen der Urheber nicht mehr zu ermitteln ist oder bei denen es aus anderen Gründen schwierig oder gar unmöglich ist, die genaue Dauer des Urheberrechtsschutzes zu bestimmen.[3]

Juristisch betrachtet, ist die Archivierung von digitalen Objekten vor allen Dingen deswegen problematisch, weil die Objekte im Normalfall für die Archivierung **kopiert** werden müssen. Für das Kopieren von Werken stellt das deutsche Urheberrecht aber bestimmte Hürden auf.

Unter bestimmten Umständen dürfen auch urheberrechtlich geschützte Werke kopiert und archiviert werden. Der einfachste Fall ist das Vorliegen einer ausdrücklichen oder konkludenten Zustimmung des Urheberrechtsinhabers. Bei Internetpublikationen ist das häufig der Fall, etwa wenn auf bestimmte Lizenzmodelle Bezug genommen wird (GNU *GPL*, Creative Commons etc.). Aus dem bloßen Einstellen von Inhalten im Internet alleine kann aber nicht auf eine konkludente Zustimmung geschlossen werden, denn aus der Tatsache, dass jemand etwas öffentlich zugänglich macht, kann nicht geschlossen werden, dass er auch damit einverstanden ist, wenn sein Angebot kopiert und dauerhaft gespeichert wird (und die Kopie womöglich seinem weiteren Zugriff entzogen ist). Zudem sind Anbieter und Urheber eines Internetangebots oft nicht identisch. Dann kann der Anbieter einem Dritten schon deswegen kein Recht zur Vervielfältigung einräumen, weil er selbst im Zweifel dieses Recht nicht hat. Anders ausgedrückt: Es ist ohne zusätzliche Zustimmung nicht erlaubt, eine interessant erscheinende Website zu Archivierungszwecken zu kopieren. Ausnahmen können sich aber ergeben, wenn zugunsten der archivierenden Institution eine spezialgesetzliche Ermächtigung besteht. Dies kann beispielsweise

2 In Einzelfällen kann es auch bei gemeinfreien Werken und digitalen Objekten, die nicht dem Urheberrecht unterliegen (z.B. wettbewerbsrechtliche) Schranken geben. Die sollen an dieser Stelle aber nicht weiter diskutiert werden. Näher dazu: Rehbinder: Urheberrecht, Rn. 103.

3 Kuhlen, Rainer: Urheberrechts-Landminen beseitigen. Bedarf nach einer Urheberrechtslösung für verwaiste Werke. http://www.kuhlen.name/Publikationen2007/verwaisteWerke-Publikation-RK0307.pdf [27.9.2007].

im Bundesarchivgesetz oder im Gesetz über die Deutsche Nationalbibliothek der Fall sein.[4]

2. Wie darf gesammelt werden?

Digitale Langzeitarchive lassen sich im Prinzip auf zweierlei Weisen füllen. Zum einen können analoge oder digitale Objekte, die sich bereits im Besitz einer archivierenden Institution befinden, ins Archiv übernommen werden. Im Regelfall setzt dies die vorherige Anfertigung einer Archivkopie oder, im Falle von analogen Objekten, deren Digitalisierung voraus. Zum anderen können auch Objekte, die sich nicht im Besitz der Institution befinden (sondern beispielsweise frei zugänglich im Internet) in das Archiv übernommen werden. Beide Wege sind nur innerhalb bestimmter rechtlicher Grenzen erlaubt. Das Problem ist auch hier jeweils, dass das Anfertigen von Vervielfältigungen nicht gemeinfreier Werke (s.o.) regelmäßig einer Zustimmung des Urheberrechtsinhabers bedarf. Es gibt jedoch wichtige Ausnahmen.

a. Anfertigung von Archivkopien

Auf den ersten Blick erscheint es naheliegend, von ohnehin vorhandenen digitalen Objekten Kopien anzufertigen, um diese dauerhaft zu archivieren. Ebenso naheliegend scheint es, analoge Objekte, die sich sowieso im Besitz der archivierenden Institution befinden, zu digitalisieren und die Digitalisate zu archivieren.

Die wichtigste Norm im Urheberrecht, die eine Anfertigung von solchen Archivkopien auch ohne Zustimmung eines Urhebers erlaubt, steht in § 53 Abs. 2 Satz 1 Nr. 2 UrhG. Demnach sind Vervielfältigungen (und darum handelt es sich bei einer Digitalisierung) gestattet, wenn die Vervielfältigung ausschließlich zur Aufnahme in ein eigenes Archiv erfolgt. Dies gilt aber nur mit wichtigen Einschränkungen:

- Die Vervielfältigung darf ausschließlich der Sicherung und internen Nutzung des vorhandenen Bestandes dienen (Archivierungszweck). Unzulässig ist hingegen die Verfolgung sonstiger Zwecke, wie etwa einer Erweiterung des eigenen Bestandes.
- Als Kopiervorlage muss ein „eigenes Werkstück" dienen. Für jede einzelne Archivierung ist dabei jeweils ein Original im Eigentum der ar-

4 Da das Urheberrechtsgesetz Bundesrecht ist, muss auch das Spezialgesetz Bundesrecht sein. Wenn also beispielsweise eine Landesbibliothek, ein Landesmuseum oder ein Landesarchiv durch Landesgesetz zur urheberrechtswidrigen Maßnahmen ermächtigt würde, wäre dies ungültig.

chivierenden Institution erforderlich, selbst dann, wenn die ansonsten identischen Kopien nur unter anderen Schlagworten abgelegt werden sollen.[5]

- Es muss sich um ein Archiv handeln, das im öffentlichen Interesse tätig ist und keinerlei wirtschaftlichen Zweck verfolgt. Gewerbliche Unternehmen, anders als beispielsweise gemeinnützige Stiftungen, sind also nicht privilegiert und dürfen ohne ausdrückliche Zustimmung der Urheberrechtsinhaber keine elektronischen Archive anlegen. Ihnen bleibt nur die analoge Archivierung, beispielsweise durch Mikroverfilmung.
- Von „Datenbankwerken" dürfen keine Archivkopien angefertigt werden (§ 53 Abs. 5 UrhG). „Datenbankwerke" sind Sammlungen von „Werken, Daten oder anderen unabhängigen Elementen, die systematisch oder methodisch angeordnet und einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sind" (§ 87a Abs. 1 UrhG)[6]. Hierzu zählen auch komplexere Webseiten.[7]
- Technische Kopierschutzverfahren dürfen nicht entfernt oder umgangen werden. Befindet sich beispielsweise eine kopiergeschützte CD-ROM im Besitz einer Gedächtnisorganisation und will diese die darauf befindlichen Daten archivieren, dann darf der Kopierschutz nicht ohne weiteres umgangen werden (§ 95a UrhG). Die Gedächtnisorganisation hat allerdings einen Anspruch darauf, dass der Rechteinhaber (z.B. der Hersteller der CD-ROM), die zur Umgehung des Schutzes erforderlichen Mittel zur Verfügung stellt, wenn die geplante Archivkopie ansonsten erlaubt ist (§ 95b UrhG). Größere Institutionen können auch mit der herstellenden Industrie pauschale Vereinbarungen treffen.[8]

b. Harvesting

Vor besondere rechtliche Probleme stellt das Harvesting von Internetangeboten, und zwar unabhängig davon, ob nach bestimmten Selektionskriterien (etwa bestimmten Suchworten) oder unspezifisch (etwa eine ganze Top-Level-Domain) gesammelt wird. Obwohl Harvesting ein gängiges Verfahren im Internet ist (vgl. etwa die Angebote von Google Cache oder archive.org), ist es nach derzeitiger Rechtslage in Deutschland nicht unproblematisch. Das Harves-

5 BGHZ 134, 250 – CB-Infobank I.

6 Die Unterscheidung des Gesetzgebers zwischen „Datenbankwerken" (§ 4 UrhG) einerseits und „Datenbanken" (§ 87a ff. UrhG) andererseits ist in diesem Fall unbeachtlich.

7 Vgl. z.B. LG Köln NJW-COR 1999, 248 L; LG Köln CR 2000, 400 – kidnet.de.

8 Vgl. die Vereinbarung zwischen dem Bundesverband der phonographischen Wirtschaft, dem Deutschen Börsenverein und der Deutschen Nationalbibliothek: http://www.ddb.de/wir/recht/vereinbarung.htm [27.9.2007].

ting ist jedenfalls dann zulässig, wenn die Zustimmung des Urhebers vorliegt (wenn beispielsweise die Betreiber einer museal interessanten Homepage einem Museum gestatten, in regelmäßigen Abständen ein automatisiertes Abbild der Homepage zu machen und dieses zu archivieren). Ohne Zustimmung des Urhebers darf keine Archivkopie angefertigt werden.
In einigen Rechtsgebieten, insbesondere den USA, kann von einer Zustimmung ausgegangen werden, wenn einer Speicherung nicht ausdrücklich widersprochen wurde und auch im Nachhinein kein Widerspruch erfolgt.[9] Nach deutscher Rechtslage reicht dies nicht aus. Die Zustimmung muss eindeutig sein. Ausnahmen, die ein Harvesting durch bestimmte Gedächtnisorganisationen gestatten, sind nur über spezielle Bundesgesetze möglich. Beispielsweise soll nach dessen amtlicher Begründung das Gesetz über die Deutsche Nationalbibliothek dieser den Einsatz von Harvesting-Verfahren ermöglichen.[10]

3. Wann und wie dürfen Archivobjekte verändert werden?

a. Migration und Emulation

Im Sinne einer langfristigen Verfügbarkeit der archivierten Objekte müssen diese gelegentlich migriert oder emuliert werden. Bei jeder Migration und, in eingeschränkterem Maße, auch bei jeder Emulation[11] kommt es zu gewissen qualitativen und/oder quantitativen Änderungen am jeweiligen Objekt. Das Wesen von Migrationen und Emulationen besteht gerade darin, die Interpretation digitaler Daten, die aufgrund ihres veralteten Formats wertlos sind, zu sichern, um sie weiterhin nutzen zu können. Diesem Ziel wird aber nur entsprochen, wenn

9 „Google Cache", „Archive.org" und vergleichbare Harvester respektieren robots.txt Dateien, über die eine Speicherung untersagt wird. Zudem werden auf Antrag des Rechteinhabers Seiten aus dem Archiv gelöscht. Zur Rechtslage in den USA vgl. das Urteil „Blake A. Field v. Google Inc. (No. 2:04-CV-0413, D.Nev)" (Online unter: http://www.linksandlaw.com/decisions-148-google-cache.htm [27.9.2007])

10 Vgl. die amtliche Begründung zu § 2 Nummer 1 des DNBG: http://www.ddb.de/wir/pdf/dnbg_begruendung_d.pdf [27.9.2007]. Ob und inwieweit das Gesetz tatsächlich den Einsatz von Harvesting-Verfahren erlaubt, muss an dieser Stelle nicht geklärt werden.

11 Es kommt dabei nicht darauf an, ob der Bitstream des ursprünglichen Objekts selbst verändert wurde, um die Abbildung auf einem neueren System zu ermöglichen. Entscheidend ist vielmehr das Erscheinungsbild für den Nutzer. In einer ganz anderen Hard- und Softwareumgebung kann im Einzelfall auch ein Objekt, dessen Daten selbst vollkommen unverändert geblieben sind, so anders erscheinen, dass von einer Umgestaltung des ursprünglichen Objekts gesprochen werden kann.

die neuen Dateien trotz etwaiger Veränderungen denselben Kern von Informationen aufweisen wie die veralteten. Dieser wesentliche Informationskern stellt sicher, dass die neue Datei durch dieselben schöpferischen Elemente geprägt sein wird wie die alte.
Entgegen gewichtigen Stimmen in der juristischen Literatur[12], handelt es sich bei den notwendigen Änderungen im Erscheinungsbild des Objekts in der Regel noch nicht um eine – zustimmungspflichtige – Bearbeitung / Umgestaltung im Sinne des § 23 UrhG, sondern um eine Vervielfältigung (§ 16 UrhG). Zum einen sind die Änderung eines Dateiformates oder das Öffnen einer Datei in einer emulierten EDV-Umgebung rein mechanische Vorgänge, die nicht von einem individuellen Schaffen desjenigen geprägt sind, der diese Vorgänge technisch umsetzt. Zum anderen kommt es bei (rechtlich unproblematischeren) Vervielfältigungen ebenfalls häufig zu kleineren Abweichungen. Solange die Vervielfältigungsstücke jedoch ohne eigene schöpferische Ausdruckskraft geblieben sind, sie noch im Schutzbereich des Originals liegen und ein übereinstimmender Gesamteindruck besteht,[13] reichen auch gewisse Detailabweichungen vom Original nicht, um von einer Bearbeitung/Umgestaltung auszugehen.
Mit anderen Worten: Soweit eine Institution das Recht hat, Kopien anzufertigen (z.B. aus dem erwähnten § 53 Abs. 2 UrhG), darf sie auch migrieren oder emulieren. Nur in den Ausnahmefällen, in denen die Migration zu einer deutlichen Abweichung vom Original führt, bedarf es einer zusätzlichen Zustimmung des Urhebers.

4. Wer darf von wo auf die archivierten Objekte zugreifen?

Der Archivbegriff der Informationswissenschaften unterscheidet sich wesentlich von dem des Urheberrechts. Während in den Informationswissenschaften auch und gerade die Erschließung und Zugänglichmachung der archivierten Materialien im Vordergrund stehen, ist der Archivbegriff in § 53 Abs. 2 UrhG deutlich enger. Hier werden ausschließlich die Sammlung, Aufbewahrung und Bestandssicherung als Archivzwecke angenommen. Ein Archiv dessen Zweck in der Benutzung durch außenstehende Dritte liegt, ist daher kein Archiv im Sinne des § 53 UrhG. Damit sind die meisten klassischen Gedächtnisorganisationen, die ihre Aufgabe in der Informationsversorgung ihrer Nutzer und weniger im Sammeln und Sichern der Bestände sehen, auf den ersten Blick von der Privilegierung des § 53 ausgenommen. Sie dürften daher ohne ausdrückliche Zustim-

12 Hoeren: Rechtsfragen zur Langzeitarchivierung, S. 7-9.

13 BGH GRUR 1988, 533, 535; Schulze-Dreier/Schulze: UrhG, § 16 Rn. 10.

mung der jeweiligen Rechteinhaber keine Vervielfältigungen anfertigen. Eine Langzeitarchivierung digitaler Daten ohne – unter praktischen Vorzeichen oft nur schwer zu erlangende – Zustimmung wäre damit *de facto* unmöglich.
Die Berechtigung, Archivkopien anzufertigen, hängt also wesentlich davon ab, ob und inwiefern außenstehende Nutzer Zugang zu den Archivmaterialien erlangen sollen. Hier sind grundsätzlich drei Varianten denkbar: rein interne Nutzung, eingeschränkte Nutzung und eine offene Nutzung.

a. Interne Nutzung

Noch verhältnismäßig unproblematisch ist eine rein interne Nutzung. Wenn Daten aus einem digitalen Archiv ausschließlich von den Mitarbeitern des Archivs im Rahmen des Archivzweckes eingesehen werden, ist dies gestattet. Schwierig wird es jedoch bereits, wenn Mitarbeiter, zum Beispiel per Download oder Computerausdruck, weitere Vervielfältigungen herstellen. Hier muss jeweils erneut geprüft werden, ob diese Vervielfältigungen auch ohne Zustimmung des Urhebers erlaubt sind (z.B. aus Gründen der wissenschaftlichen Forschung – § 53 Abs. 2 S. 1 Nr. 1 UrhG).

b. Nutzung durch einen begrenzten Nutzerkreis

Der neu eingefügte § 52b UrhG gestattet es öffentlichen Bibliotheken, Museen und Archiven, ihren Bestand an eigens dafür eingerichteten elektronischen Leseplätzen zugänglich zu machen. Analoge Bestände dürfen zu diesem Zweck digitalisiert werden und bereits vorhandene Archivdigitalisate in den gesteckten Grenzen öffentlich zugänglich gemacht werden.
§ 52b UrhG enthält aber auch wichtige Beschränkungen, die es zu beachten gilt.

- Privilegiert werden nur nichtkommerzielle öffentliche Bibliotheken, Museen und Archive. Nicht-öffentliche Bibliotheken, wie Schul-, Forschungseinrichtungs- oder Institutsbibliotheken oder gewerbliche Archive dürfen sich nicht auf § 52b UrhG berufen.
- Die Anzahl der erlaubten Zugriffe an den eingerichteten Leseplätzen richtet sich grundsätzlich nach der Zahl des in der Gedächtnisorganisation vorhandenen Bestandes.
- Vertragliche Vereinbarungen (etwa Datenbanklizenzen) gehen vor. Wenn die Nutzung durch Dritte vertraglich ausgeschlossen worden ist, kann dies nicht unter Berufung auf § 52b UrhG umgangen werden.

Ähnlich wie bei einer internen Nutzung ist zu entscheiden, ob und wann Nutzer Downloaden oder Ausdrucken dürfen (s.o.).
Wenn aus einem der genannten Gründe § 52b UrhG nicht greift (etwa, weil es

sich bei der archivierenden Institution um eine nicht-öffentliche Forschungsbibliothek handelt), bleibt die Frage, inwieweit die Institution ihren Nutzern Zugang zu den archivierten Materialien gewähren darf. Dies ist in bestimmten Fällen möglich. Beispielsweise ist die Zugänglichmachung von kleinen Teilen von Werken, kleineren Werken und einzelnen Zeitungs- oder Zeitschriftenbeiträgen durch (eng) abgrenzte Personengruppen, z.B. einzelnen Forscherteams oder den Teilnehmern eines Universitätsseminars, erlaubt, soweit die Nutzung dabei zum Zwecke der wissenschaftlichen Forschung oder zu Unterrichtszwecken (§ 52a UrhG) erfolgt.[14]

c. Offene externe Nutzung

Es gehört zum Charme der neuen Medien und insbesondere des Internets, dass sie im Prinzip einen weltweiten Zugriff ermöglichen. Der Gesetzgeber hat aber die Entscheidung darüber, ob ein digitales Objekt einer breiten Öffentlichkeit zugänglich gemacht werden soll, alleine dem Urheber übertragen. Ohne Zustimmung des Urhebers darf also keine Gedächtnisorganisation urheberrechtlich geschütztes Material ortsungebunden öffentlich zugänglich machen.

5. Wer haftet für die Inhalte?

Wenn eine Gedächtnisorganisation in großem Umfang digitale Objekte der mehr oder weniger breiten Öffentlichkeit anbietet, besteht die Gefahr, dass einige der Objekte durch ihren Inhalt gegen Rechtsnormen verstoßen. Volksverhetzende oder pornografische Inhalte lassen sich durch entsprechende Filtersoftware und im Idealfall eine intellektuelle Sichtung des Materials noch relativ leicht erkennen. Oft ist es aber nahezu unmöglich, ehrverletzende Behauptungen oder Marken- und Patentverletzungen zu identifizieren. Es ist also eine wichtige Frage, welche Sorgfaltspflichten eine Gedächtnisorganisation zu beachten hat, die ihre digitalen Archivalien öffentlich zugänglich machen will.
Leider ist hier so vieles vom konkreten Einzelfall abhängig, dass es sich nicht mehr wirklich sinnvoll in einer kurzen Zusammenfassung darstellen lässt. Eine ausführlichere Darstellung würde aber den hier vorgegebenen Rahmen sprengen. Nur ganz allgemein kann Folgendes gesagt werden:
Die in diesem Bereich wichtigsten Normen stehen in den §§ 7 - 10 Telemediengesetz (TMG). Danach ist zu unterscheiden, ob es sich bei den veröffentlichten Inhalten um eigene oder fremde handelt. Eine straf- und zivilrechtliche Verant-

14 Das gilt auch für den Zugang zu Vervielfältigungsstücken, die zu Archivzwecken angefertigt worden sind (§ 53 Abs. 2 S. 1 Nr. 2 UrhG).

wortung für die Richtigkeit und Rechtmäßigkeit der Inhalte trifft die anbietende Organisation nur im ersten Fall. Ob die Inhalte im Einzelfall der Organisation als eigene zugerechnet werden, richtet sich dabei nicht nach Herkunft oder Eigentum der Objekte, sondern nach der Sicht der Nutzer.[15] Nur wenn ein Nutzer aus den Gesamtumständen eindeutig erkennen konnte, dass es sich bei dem Angebot nicht um ein eigenes Informationsangebot der betreffenden Organisation handelt, ist die Haftung eingeschränkt. Eine Gedächtnisorganisation, die fremde Daten allgemein zugänglich macht, sollte daher darauf achten, dass die „fremden" Angebote im Layout hinreichend deutlich von den eigenen abgegrenzt sind. Außerdem sollte deutlich darauf hingewiesen werden, dass sich die Gedächtnisorganisation nicht mit den Inhalten der angebotenen Publikationen oder verlinkten Seiten identifiziert und eine Haftung für diese Inhalte ausgeschlossen ist. Hiermit stellt sie klar, dass sie lediglich dann zur Haftung herangezogen werden kann, wenn sie falsche oder rechtswidrige Inhalte trotz Kenntnis oder Evidenz nicht beseitigt.

Auch wenn deutlich gemacht wurde, dass die zugänglich gemachten Inhalte keine eigenen sind, müssen bestimmte Sorgfaltspflichten beachtet werden. Vor allen Dingen muss bei Bekanntwerden einer Rechtsverletzung der Zugang unverzüglich gesperrt werden (§ 7 Abs. 2 TMG). Eine weitere Speicherung des Objektes bleibt aber – von wenigen Ausnahmen abgesehen – möglich, denn nur die Zugänglichmachung muss unterbunden werden.

Literatur

Dreier, Thomas / Schulze, Gernot: Urheberrechtsgesetz: Urheberrechtswahrnehmungsgesetz, Kunsturhebergesetz; Kommentar. München: Beck, 2004

Dreyer, Gunda / Kotthoff, Jost / Meckel, Astrid: Heidelberger Kommentar zum Urheberrechtsgesetz. Heidelberg: Müller, 2004

15 Das ist im Falle von Gedächtnisorganisationen schwierig, handelt es sich doch um Material aus eigenen Archiven. In einem bestimmten Sinne ist also auch das angebotene Archivmaterial „eigen" und wird insbesondere nicht „für einen Nutzer" (§ 10 TMG) gespeichert. Trotzdem ist es klar ersichtlich und ergibt sich meist auch aus dem (oft gesetzlichen) Auftrag der Gedächtnisorganisation, dass sie sich die angebotenen Inhalte nicht zu Eigen machen will und kann. Eine Haftung als Content-Provider wäre daher unbillig. Vielmehr ist § 10 TMG zugunsten der jeweiligen Gedächtnisorganisation analog anzuwenden, wenn die Abgrenzung der Inhalte, die im engeren Sinne „eigen" sind und denjenigen, die als „fremde" zur Verfügung gestellt werden, hinreichend deutlich ist.

Forgó, Nikolaus: Urheberrecht in digitalisierter Forschung und Lehre. Hannover: Jur. Fakultät, 2006

Goebel, Jürgen W. / Scheller, Jürgen: Digitale Langzeitarchivierung und Recht; nestor-Materialien 01: urn:nbn:de:0008-20040916022

Hoeren, Thomas: Informationsrecht: http://www.uni-muenster.de/Jura.itm/hoeren/

material/Skript/skript_maerz2007.pdf [letzter Zugriff: 27.9.2007]

Hoeren, Thomas: Rechtsfragen zur Langzeitarchivierung (LZA) und zum Anbieten von digitalen Dokumenten durch Archivbibliotheken unter besonderer Berücksichtigung von Online-Hochschulschriften: urn:nbn:de:0008-20050305016

Kuhlen, Rainer: Urheberrechts-Landminen beseitigen: Bedarf nach einer Urheberrechtslösung für verwaiste Werke: http://www.kuhlen.name/Publikationen2007/verwaisteWerke-Publikation-RK0307.pdf [letzter Zugriff: 27.9.2007]

Ott, Stephan: Der Google Cache – Eine milliardenfache Urheberrechtsverletzung? In: MIR 2007, Dok.195: http://medien-internet-und-recht.de/volltext.php?mir_dok_id=697

Rehbinder, Manfred: Urheberrecht: Ein Studienbuch. 14. Auflage, München: Beck, 2006

Schack, Haimo: Dürfen öffentliche Einrichtungen elektronische Archive anlegen? In: AfP – Zeitschrift für Medien- und Kommunikationsrecht 1/2003, S. 1-8

3 State of the Art

LZA-Aktivitäten in Deutschland aus dem Blickwinkel von nestor

Dr. Mathias Jehn, Sabine Schrimpf

Die Situation in Deutschland

Bibliotheken, Archive und Museen sind das wissenschaftliche, juristisch-administrative und kulturelle Gedächtnis einer Stadt, eines Landes, einer Nation. Sie sind Orte der Forschung und Wissensvermittlung, des Lernens und der Anschauung. Sie tragen die Verantwortung für die Erhaltung physisch vorhandener Originale ebenso wie für die langfristige Nutzbarkeit digitaler Informationen bzw. nachträglich angefertigter Digitalisate von anderen Kulturmedien. Gerade elektronische Publikationen oder, weiter gefasst, digitale Ressourcen nehmen in den meisten deutschen Einrichtungen einen stetig wachsenden Stellenwert ein und beeinflussen nachhaltig den Auftrag von Gedächtnisorganisationen. Die rasante Entwicklung auf diesem Gebiet stellt neue Anforderungen

hinsichtlich der dauerhaften Bewahrung und Zugänglichkeit dieser digitalen Objekte: So muss das digital publizierte Wissen auch unter den Bedingungen eines ständig stattfindenden Technologiewandels langfristig verfügbar gehalten werden, da der wissenschaftliche und technische Fortschritt eine regelmäßige Neubewertung älterer Wissensstände erfordert. Der digitalen Langzeitarchivierung kommt hierbei eine Schlüsselrolle zu. Letztlich stellt sie eine wesentliche Bedingung für die Konkurrenzfähigkeit des Bildungs- und Wissenschaftssystems und damit mittelbar auch für die wirtschaftliche Leistungsfähigkeit eines jeweiligen Landes dar.

Die digitale Langzeitsicherung erweitert das Aufgabenspektrum der archivierenden Institutionen, sodass neue organisatorische und technische Anstrengungen zur Sicherung und langfristigen Nutzbarkeit digitaler Objekte erforderlich sind. Ein Archiv, das sich erst bei anstehenden Lieferungen des elektronischen Schriftguts Gedanken über dessen Übernahme, Erschließung und die dauerhafte Speicherung macht, wird an der Komplexität der Aufgabe scheitern. Die dauerhafte Lesbarkeit von elektronischen Medien ist insbesondere durch den schnellen technischen Wandel von Datenträgern und -formaten sowie durch die permanente Veränderung und Weiterentwicklung der für die Nutzung notwendigen Anwendungsprogramme gefährdet. Die Arbeit, die im Bereich der physischen Datenträger geleistet wurde, vorgegeben durch gesetzliche Sammelaufträge oder Archivgesetze, hat deutlich werden lassen, dass sowohl für solch große Bereiche der Netzpublikationen wie ebooks, e-Journals, elektronische Hochschulschriften oder thematische Websites (bzw. Online-Ressourcen) gemeinsame und tragfähige Langzeitarchivierungsstrategien bislang noch fehlten. Dazu kommt, dass die Aufgaben sich in eine Vielzahl von Teilaspekten gliedern und daraus resultierenden Teilaufgaben von einer Institution allein nicht zu leisten sind. Neben den Bibliotheken werden auch die Archive in Zukunft mit einer wachsenden Zahl von Abgaben elektronischen Schriftguts rechnen müssen. Dieses Schriftgut aus den Behörden wird von Anfang an elektronisch („digital born“) erstellt und voraussichtlich die volle Bandbreite an Formen digitaler Unterlagen umfassen.

In Deutschland wurde das Thema zum ersten Mal 1995 in einem Positionspapier „Elektronische Publikationen“ der Deutschen Forschungsgemeinschaft (DFG) aufgegriffen und als Aufgabenbereich der Virtuellen Fachbibliotheken benannt. In Anbetracht sowohl des Umfangs der Aufgabe als auch der föderalen Struktur Deutschlands mit der Verantwortlichkeit seiner Bundesländer für Wissenschaft und Kultur, war es folgerichtig, dass der Ansatz zu einer

erfolgreichen Lösung dieser Probleme nur ein kooperativer sein konnte. Aus der gemeinsamen Arbeit an konzeptionellen Fragen der künftigen Entwicklung digitaler Bibliotheken im Rahmen des vom Bundesministeriums für Wissenschaft und Forschung (BMBF) getragenen Projektes „digital library konzepte“ ist eine Initiativgruppe Langzeitarchivierung hervorgegangen, deren Arbeitsplan im Rahmen einer 6-monatigen Folgeprojekts im Jahre 2002 auf zwei Workshops ausgewählten Experten des Informationswesens zur Diskussion gestellt wurden. Diese „Initialzündung“ für eine kooperative Lösung der Langzeitarchivierung digitaler Ressourcen resultierte in einem Papier mit Abschlussempfehlungen für zentrale Komponenten einer kooperativen digitalen Langzeiterhaltungsstrategie für Deutschland. Seit dem Jahr 2003 besteht mit dem BMBF-geförderten Projekt *nestor* ein nationales Kompetenznetzwerk zur Langzeitarchivierung und Langzeitverfügbarkeit digitaler Objekte, das als einziges seiner Art die in Deutschland identifizierbaren Kompetenzen bündelt und die Kontakte zu entsprechenden Initiativen und Fachgruppen koordiniert.[1] Mit der Einrichtung von *nestor* sollte gemeinsam den Defiziten bei der Langzeitarchivierung – unter Einbeziehung der „Produzenten“ digitaler Ressourcen, d. h. Verlage, Universitäten, Forschungseinrichtungen, Behörden, Wissenschaftler sowie technischer Dienstleister wie Rechen-, Daten- und Medienzentren und Großdatenbankbetreiber – begegnet werden. Die gemeinsame Fragestellung betrifft die dauerhafte Erhaltung sowohl genuin digitaler Objekte als auch retrodigitalisierter Ressourcen sowie die nachhaltige Verfügbarkeit dieser Informationen für spätere Generationen.

Mittlerweile verteilen sich in *nestor* die notwendigen Fachkompetenzen für den Aufgabenkomplex „Langzeitarchivierung digitaler Ressourcen“ über ein breites Spektrum von Personen, die in vielen Institutionen, Organisationen und Wirtschaftsunternehmen tätig sind. *nestor* bringt so die Experten der Langzeitarchivierung und aktive Projektnehmer zusammen und fördert den Austausch von Informationen, die Entwicklung von Standards sowie die Nutzung von Synergieeffekten. Alle Sparten der Gedächtnisinstitutionen werden bei der Herausforderung unterstützt, die Bewahrung und Verfügbarkeit aller digitalen Ressourcen selbst zu gewährleisten, die Bewahrung und Verfügbarkeit der wichtigsten Objekte jedes Fachgebiets zu organisieren sowie schließlich die Bewahrung und Verfügbarkeit digitaler Archivalien garantieren zu können.

---

1 nestor ist das Akronym der englischen Übersetzung des Projekttitels: „Network of Expertise in long-term storage and availability of digital Resources in Germany“. Siehe: http://www.langzeitarchivierung.de.

Für Bibliotheken, Archive und Museen ist mit der Einrichtung von *nestor* ein wichtiger Schritt für die verteilte Übernahme konkreter Aufgaben und Absprachen in Deutschland getan. Ein zentrales Aufgabenfeld des Netwerks ist beispielsweise die Sicherung der Authentizität (im Sinne der Vertrauenswürdigkeit) des archivierten Dokuments. Im Prozess der Planung von Erhaltungsstrategien sind u.a. drei wichtige Arbeitsschritte zu vollziehen:

1. Da ein nationaler Alleingang in der globalen Informationsgesellschaft ein sicherer Misserfolgsfaktor wäre, ist es erstens wichtig, eine Bestandsaufnahme, Analyse und Auswertung der internationalen Entwicklungen vorzunehmen und zu prüfen, welche der bereits existierenden Lösungsvorschläge der deutschen Situation angemessen sein könnten.
2. Die Entwicklung von Norm-Standards ist unbedingt erforderlich. Diese sollten in Übereinstimmung mit den sich aktuell im internationalen Rahmen abzeichnenden Standardisierungsinitiativen erarbeitet werden.
3. Der Aufbau einer dezentralen und kooperativen Infrastruktur für die Archivierung digitaler Dokumente in Deutschland, die nicht nur Zuständigkeiten klar definiert sondern auch effektive und effiziente Kooperationsstrukturen etabliert, ist notwendig.

Zur Umsetzung dieser Ziele müssen weitere finanzielle Mittel zur Verfügung gestellt werden, weil mit der Langzeitarchivierung und -verfügbarkeit digitaler Objekte völlig unterschiedliche Bereiche betroffen sind.[2] Sobald einmal mit der Langzeitarchivierung begonnen wird, muss die langfristige Finanzierung gewährleistet sein. Zwar ist heute immer noch unklar, wie sich die Kosten in der Zukunft entwickeln werden, jedoch ist es sicher, dass einerseits große Geldsummen für den Aufbau und Betrieb von Langzeitarchivierungssystemen benötigt werden, andererseits der finanzielle Spielraum für den öffentlich-rechtlichen Bereich begrenzt sein wird. Es sind daher Strategien nötig, wie Gedächtnisorganisationen mit den begrenzten Mitteln die besten Ergebnisse erzielen können.

Kurzer Überblick über die Langzeitarchivierungssysteme und -

2 Ein wichtiges Ergebnis der ersten Projektphase von 2003 bis 2006 war die Verabschiedung gemeinsamer Richtlinien: nestor hat in einem „Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland“ die notwendigen Anstrengungen von politischen Entscheidungsträgern, Urhebern, Verlegern, Hard- und Softwareherstellern sowie kulturellen und wissenschaftlichen Gedächtnisorganisationen zusammengestellt, um die Rahmenbedingungen einer nationalen Langzeitarchivierungs-Policy abzustecken. Siehe: http://www.langzeitarchivierung.de/downloads/memo2006.pdf.

projekte

In Deutschland gibt es schon einige Institutionen, die mit der digitalen Langzeitarchivierung begonnen haben. Auf Grund der komplexen und innovativen Herausforderungen, die mit dem Thema digitale Langzeitarchivierung verbunden sind, geschieht dies meist im Projektverbund.

Mit kopal („kooperativer Aufbau eines Langzeitarchivs digitaler Informationen") haben die Deutsche Nationalbibliothek in Kooperation mit der Niedersächsischen Staats- und Universitätsbibliothek (SUB) Göttingen, der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und IBM Deutschland kooperativ eine technische Lösung für die Bewahrung und langfristige Verfügbarkeit digitaler Dokumente erarbeitet.[3] Seit der Aufnahme des Produktivbetriebs im August 2006 hat kopal umfangreiche digitale Bestände von der Deutschen Nationalbibliothek und der SUB Göttingen in das Archivsystem eingespielt. Die beteiligten Institutionen konnten dabei wertvolle Erfahrungen für den Betrieb und die weitere Entwicklung des Archivsystems sammeln. Das kopal-Langzeitarchiv ist nach dem Projektende im Juni 2007 bei der Deutschen Nationalbibliothek und der (SUB) Göttingen, in den Routinebetrieb gegangen. Anlässlich des Abschlussworkshops „kopal goes live" am 13. Juni 2007 wurde ein Memorandum unterzeichnet, in dem sich die Partner langfristig den Aufgaben der Langzeitarchivierung verpflichten und einen Rahmen für die weitere Zusammenarbeit gesetzt haben.

Daneben wurde mit dem Pilotsystem „Bibliothekarisches Archivierungs- und Bereitstellungssystem" wurde ein weiteres Archivsystem an der Bayerischen Staatsbibliothek München in Zusammenarbeit mit dem Leibniz Rechenzentrum entwickelt.[4] Ziel des von der DFG geförderten Kooperationsprojektes war der Aufbau einer organisatorischen und technischen Infrastruktur für die Langzeitarchivierung und Bereitstellung von Netzpublikationen aus dem breiten Spektrum der Bayerischen Staatsbibliothek als Universal-, Landes- und SSG-Bibliothek sowie als Digitalisierungszentrum. Im Nachfolgeprojekt BABS II soll das Pilotsystem zu einem vertrauenswürdigen digitalen Langzeitarchivs als Teil kooperativer Strukturen und Evaluierung gemäß dem *nestor*-Kriterienkatalog ausgebaut werden. Evaluierbarkeit und Test der Skalierbarkeit des Gesamtsystems sollen einen langfristigen Betrieb mit Wachstumspotential gewährleisten.
Mit edoweb in Rheinland-Pfalz, BOA in Baden-Württemberg und Saardok im

3 http://kopal.langzeitarchivierung.de/

4 http://www.babs-muenchen.de/index.html?pcontent=startseite

Saarland liegen kooperativ entwickelt und betriebene technische Plattformen für die Sammlung, Erschließung und langfristige Verfügbarkeit von regionalen elektronischen Pflichtexemplaren vor.[5]

Neben der Entwicklung kompletter Archivsystem-Lösungen befassen sich zahlreiche Institutionen in unterschiedlichen Projekten mit weiteren Aspekten der digitalen Langzeitarchivierung. *nestor* bündelt alle derartigen Projekte in Deutschland, im deutschsprachigen Raum sowie die mit Beteiligung deutscher Partner auf der *nestor*-Homepage. Das Themenspektrum der aufgeführten Projekte reicht von den hier beispielhaft vorgestellten Archivsystemen über die Strategiebildung hinsichtlich Langzeitarchivierung bis zur konkreten Entwicklung von Langzeitarchivierungswerkzeugen.

Neben diesen Beispielen aus Deutschland liegen auch einige gute Beispiele für erfolgreiche internationale Kooperationsprojekte im Bereich der Langzeitarchivierung vor. Im Bereich der technologischen Forschung sind die von der EU geförderten Forschungsprojekte PLANETS und CASPAR wichtige Einrichtungen,[6] etwa bei der Implementierung des Open Archival Information System, kurz OAIS-Modell.[7] Durch die Abgrenzung und eindeutige Benennung von Funktionsmodulen, Schnittstellen und Typen von Informationsobjekten ist es gelungen, eine einheitliche Sprache und eine über die Grenzen der Anwendergemeinschaften Archive, Datenzentren und Bibliotheken hinweg geltende allgemeine Sicht auf die Kernfunktionen eines digitalen Archivs zu schaffen. Gerade durch diese Allgemeingültigkeit ist der Abstraktionsgrad des Modells relativ hoch. Das Open Archival Information System beschreibt ein Informationsnetzwerk, das den Archivar und den Nutzer als Hauptkomponenten des digitalen Archivs versteht.

Auch für den Bereich der Zertifizierung von Archiven liegen bereits Ergebnisse vor, wie beispielsweise die TRAC Checkliste oder der *nestor* „Kriterienkatalog Vertrauenswürdige Archive".[8] Die hier veröffentlichten Kriterien beschreiben die organisatorischen und technischen Voraussetzungen eines digitalen Lang-

5 http://www.lbz-rlp.de/cms/rheinische-landesbibliothek/digitale-angebote/edoweb/, http://www.boa-bw.de/, http://saardok.sulb.uni-saarland.de/

6 Siehe: http://www.planets-project.eu/; http://www.casparpreserves.eu/.

7 Das als ISO 14721 verabschiedete Referenzmodell „Open Archival Information System – OAIS" ist abgedruckt in: http://public.ccsds.org/publications/archive/650x0b1.pdf.

8 Die Kriterienkataloge sind hinterlegt in: http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91.

zeitarchivs und sind auf eine Reihe digitaler Repositorien und Archive anwendbar, von universitären Repositorien bis hin zu großen Datenarchiven; von Nationalbibliotheken bis hin zu digitalen Archivierungsdiensten Dritter. Anhand der Kriterienkataloge kann die Vertrauenswürdigkeit digitaler Langzeitarchive nun geprüft und bewertet werden. Darüber hinaus beteiligen sich die Partner von *nestor* aktiv auch an europäischen Initiativen und Projekten, beispielhaft können hier DRIVER und DPE genannt werden.[9] Die Anbindung der eigenen Überlegungen an die Förderlinien der Europäischen Kommission ist wichtiger Bestandteil der Arbeit. Über Europa hinaus bestehen enge Bindungen z.B. an die frühere amerikanische Research Libraries Group und die australische Nationalbibliothek, gemeinsam mit außereuropäischen Partnern in den USA und Asien wird einmal jährlich eine internationale Konferenz organisiert (IPRES).[10]

9 Siehe: http://www.driver-repository.eu/ und http://www.digitalpreservationeurope.eu/.

10 Siehe: http://rdd.sub.uni-goettingen.de/conferences/ipres/ipres-en.html.

3.1 Bibliotheken

Mathias Jehn

Für die Bibliotheken gehört der Umgang mit elektronischen Ressourcen angesichts der sich gegenwärtig vollziehenden Veränderungen in der Informationsgesellschaft zu den größten Herausforderungen des 21. Jahrhunderts. Zwar ist die jeweilige Sichtweise auf digitale Informationen je nach Bibliothekstyp und -aufgabe traditionell sehr unterschiedlich, jedoch hat in den letzten Jahren ein Prozess intensiven Nachdenkens darüber eingesetzt, welche gemeinsamen Wege beschritten werden müssen, um dem bibliothekarischen Auftrag auch in Zukunft gerecht zu werden. Ein entscheidender Mangel konnte bis heute noch nicht behoben werden: Die Frage nach den Möglichkeiten und Bedingungen der zuverlässigen Archivierung elektronischer Ressourcen ist noch weitgehend unbeantwortet. Dies gilt sowohl für die Sicherung der Datenspeicherung (Trägermedium) als auch den zukünftigen Zugriff auf die in ihnen enthaltenen Informationen (Datenformate) und deren dauerhafte Nutzbarkeit (Erschließung und Bereitstellung). Alle Bibliotheken sind sich darüber einig, dass unter dem wachsenden Druck betriebswirtschaftlichen Denkens keine Institution allein alle digitalen Ressourcen dauerhaft archivieren kann, sondern dass geeignete nationale Kooperations- und Austauschmodelle greifen müssen. In diesem Kontext stehen die Themenfelder „Netzpublikationen", „Langzeitspeicher" und „nationales Vorgehen" im Zentrum der aktuellen Diskussion:

1. Erweiterter Sammelauftrag:

Seit der Mitte der 1990er Jahre nimmt die Bedeutung originär digitaler Publikationen stetig zu. Zahlreiche Verlage veröffentlichen wissenschaftliche Zeitschriften - besonders im naturwissenschaftlichen Bereich - auch oder ausschließlich in digitaler Form. Die zunehmende Bedeutung von Netzpublikationen erweitert das Aufgabenspektrum der Bibliotheken und befördert die organisatorischen und technischen Anstrengungen zur Sicherung und langfristigen Nutzbarkeit digitaler Objekte. Auf Empfehlung der Kultusministerkonferenz (KMK) wird von den Universitäten seit 1998 zunehmend die Veröffentlichung von Promotions- und Habilitationsarbeiten in digitaler Form akzeptiert. Pflichtexemplar- und Sondersammelgebietsbibliotheken haben in den vergangenen Jahren Kompetenzen bei der Sammlung und Bearbeitung digitaler Medien aufgebaut. Im Juni 2006 wurde das Gesetz über die Deutsche Nationalbibliothek verabschiedet; ab sofort sind elektronische Veröffentlichungen in die Regelungen über eine

nationale Sammlung und Verzeichnung einbezogen. Nach der Novellierung des Bundesgesetzes sollten die Novellierungen der einschlägigen Ländergesetze baldmöglichst folgen. Das so genannte „Drei-Varianten-Vorgehen" bietet hierbei eine Möglichkeit für das Sammeln elektronischer Publikationen. Darunter versteht man: 1. Direkte Kooperation mit den Ablieferern oder Kooperation mit aggregierenden Partnern wie regionalen Pflichtexemplarbibliotheken oder zentralen Fachbibliotheken hinsichtlich der Sammlung einzeln identifizierbarer Online-Publikationen. 2. Implementierung einer generell nutzbaren Schnittstelle auf der Website für die Ablieferung einzeln identifizierbarer Netzpublikationen in standardisierten Verfahren. 3. Erprobung von Harvesting-Methoden für die Sammlung bzw. den Abruf definierter Domainbereiche.

2. Aufbau eines Langzeitspeichers:

Die Sammlung der Netzpublikationen macht den Aufbau gewaltiger Datenspeicher erforderlich. Dies setzt neue Formen der Zusammenarbeit in Deutschland voraus. Allein die bloße Datenspeicherung genügt nicht; große Datenmengen müssen verwaltet werden, um adressierbar zu bleiben. Zudem müssen Prozesse entwickelt werden, die den „Import" neuer Daten in den Datenspeicher regeln. Darüber hinaus muss für die künftige Migration, Emulation oder Konversion der Daten zum Zweck der Langzeitarchivierung Vorsorge getroffen werden. Die Nutzbarkeit sollte gewährleistet sein, auch wenn Hard- und Softwareumgebungen und Benutzungstools technisch veralten und eine weitere Nutzbarkeit der ursprünglichen Form verhindern. All diese Fragen werden seit 2004 von der Deutschen Nationalbibliothek zusammen mit den Partnern Staats- und Universitätsbibliothek Göttingen, IBM und Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen im Projekt kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen: http://kopal.langzeitarchivierung.de/) bearbeitet. Eine erste Implementierungsstufe wurde im Frühjahr 2006 fertig gestellt. Zur dauerhaften Adressierung der Online-Objekte vergibt die Deutsche Nationalbibliothek persistente Identifikatoren in Form eines URN (Uniform Resource Name), der anders als eine Web-URL dauerhaft adressierbar und damit zitierbar bleibt.

3. Errichtung eines kooperativen Netzwerks:

Die notwendige Steuerung, Koordination, Forschung und Entwicklung für eine leistungsfähige Langzeitarchivierung fand in Deutschland in der Vergangenheit nur in geringem Umfang statt. Aus diesem Grund hat sich im Jahr 2003 mit dem Projekt *nestor* (Network of Expertise in long-term Storage and availability

of digital Ressources in Germany) erstmals ein nationales Kompetenznetzwerk gebildet, um den immer spürbarer werdenden Defiziten bei der Langzeitarchivierung gemeinsam zu begegnen. Die Partner in dem bis 2009 genehmigten Projekt sind die Deutsche Nationalbibliothek, die Staats- und Universitätsbibliothek Göttingen, die Bayerische Staatsbibliothek München, die Humboldt-Universität Berlin, das Bundesarchiv, die Fernuniversität Hagen und das Institut für Museumsforschung der Stiftung Preußischer Kulturbesitz in Berlin. Die wesentlichen Aufgaben sind: Identifikation von Arbeitsgruppen, Institutionen, Projekten, Experten im deutschsprachigen Raum, die inhaltlich zur Ausfüllung des Kompetenznetzwerkes beitragen können, Aufbau der intensiv genutzten Plattform des Kompetenznetzwerks http://www.langzeitarchivierung.de zu allen Fragestellungen der Langzeitarchivierung digitaler Ressourcen, Bewusstseinsbildung bei Bibliotheken, Archiven und Museen für die Fragestellungen der Langzeitarchivierung und für die Parallelität der Themenstellungen in den drei Communities, sowie die Durchführung von Workshops und Seminaren zu unterschiedlichen Aspekten der Langzeitarchivierung. Die Anbindung der Aktivitäten an die Förderlinien der Europäischen Kommission und die Zusammenarbeit mit außereuropäischen Institutionen sind wesentlicher Bestandteil der Arbeit. Darüber hinaus hat *nestor* in einem „Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland" die notwendigen Anstrengungen von politischen Entscheidungsträgern, Urhebern, Verlegern, Hard- und Softwareherstellern sowie kulturellen und wissenschaftlichen Gedächtnisorganisationen zusammengestellt, für die die Rahmenbedingungen in einer nationalen Langzeitarchivierungs-Policy gesichert werden müssen.

Eine wesentliche Vorbedingung für die Etablierung einer Archivierungsstruktur für elektronische Ressourcen in Deutschland ist die Stärkung der öffentlichen Bewusstseinsbildung für die Relevanz der Langzeitarchivierung elektronischer Ressourcen. Derzeit kommen die entscheidenden Entwicklungen auf diesem Gebiet vor allem aus dem angloamerikanischen Raum (USA, England, Australien). Um in Zukunft die Anschlussfähigkeit der Archivierungsaktivitäten an diese Entwicklungen zu gewährleisten und diese vor dem Hintergrund der spezifischen bibliothekarischen Bedürfnisse und Gegebenheiten der deutschen Informationslandschaft mitzugestalten, wird eine intensivere Kooperation und eine noch stärkere Partizipation der Bibliotheken an diesen Initiativen notwendig sein.

3.2 Archive

Christian Keitel

Die digitale Revolution fordert die klassischen Archive in zwei Bereichen heraus: Zum einen bedürfen die übernommenen Objekte ständiger Aufmerksamkeit und Pflege; es genügt nicht mehr, sie in einem Regal abzulegen und über Findbücher nachweisbar zu halten. Zum anderen müssen die Archive bereits vor dem Zeitpunkt der Bewertung aktiv werden, um ihren Aufgaben auch künftig nachkommen zu können. Während in den angelsächschen Ländern die Archive seit jeher auch für die Schriftgutverwaltung der abgebenden Stellen (Behörden, Unternehmen...) zuständig sind, ist die Aufgabe des Recordsmanagements für die deutschen Archive neu.

Der Lebenslauf (Lifecycle) eines digitalen Objekts kann aus Sicht des Archivs in mehrere Phasen eingeteilt werden.

1.) Systemeinführung:

Bei der Einführung eines neuen IT-Systems in der abgebenden Stelle sollte das Archiv beteiligt werden, um wenigstens die Anbietung und den Export der im System zu produzierenden Unterlagen zu gewährleisten. Neben der Definition von Schnittstellen ist dabei über geeignete Formate und die Ausgestaltung von Löschroutinen zu sprechen. Bei einem weitergehenden Anspruch kann das Archiv auch versuchen, in der Behörde auf eine authentische und integre Schriftgutverwaltung hinzuwirken. Als Standards im Bereich der Schriftgutverwaltung können genannt werden: DOMEA (Deutschland), GEVER (Schweiz), ELAK (Österreich), NOARK (Norwegen), MoReq (EU, angelsächisch geprägt) und die ISO 15489. In Australien soll sich jedes in der Behörde entstehendes Dokument über eine spezielle Nummer eindeutig dieser Behörde zuweisen lassen (AGLS). Ebenfalls sehr weit ausgearbeitet ist das VERS-Konzept aus der australischen Provinz Victoria.

2.) Bewertung:

Seit jeher können Archive nur einen Bruchteil der in den abgebenden Stellen verwahrten Unterlagen übernehmen. Die Auswahl der archivwürdigen digitalen Unterlagen weicht teilweise von der archivischen Bewertung papierner Unterlagen ab. Gemein ist beiden Prozessen der Versuch, vielfältig interpretierbare

aussagekräftige Unterlagen zu ermitteln. Dienstreiseanträge werden auch nicht dadurch archivwürdig, wenn sie in digitaler Form vorliegen. Andererseits ermöglichen digitale Unterlagen neue Formen der Informationssuche und -aggregierung. Es kann daher sinnvoll sein, in manchen Bereichen ganze Datenbanken zu übernehmen, aus denen bisher mangels Auswertbarkeit nur wenige oder keine Papierakten ins Archiv übernommen wurden. Die Diskussion über geeignete Bewertungsmodelle und -verfahren wird noch einige Jahre in Anspruch nehmen.

3.) Übernahme und Aufbereitung:

Abhängig von den bei der Systemeinführung erfolgten Absprachen bekommen die Archive im günstigsten Fall sämtliche Daten in archivfähiger Form angeboten, im schlechtesten müssen sie sich selbst um den Export und die spätere Umwandlung in taugliche Formate sowie deren Beschreibung bemühen. Die meisten Archive setzen auf das Migrationskonzept, benötigen also eine entsprechend aufwändige Aufbereitung der Daten. In zunehmendem Maß stehen dabei kleine Tools zur Verfügung, die v.a. von angelsächischen Archiven als Open Source Software veröffentlicht werden, z.B. DROID (National Archives, Kew) und XENA (National Archives of Australia).

4.) Archivierung:

Ende des letzten Jahrhunderts wurde im angelsächischen Raum das Konzept der „postcustocial option“ diskutiert. Danach sollten die datenerzeugenden Stellen diese bei festgestellter Archivwürdigkeit unbefristet aufbewahren. Den Archiven würde dann die Aufgabe der Bewertung und die Kontrolle über die Speicherung und Zugänglichkeit der Daten zufallen. Dieses Konzept wird seit einigen Jahren nicht mehr diskutiert, mit dem australischem Nationalarchiv hat sich 2000 auch ein ehemaliger Fürsprecher wieder der klassischen Übernahme und Archivierung zugewandt. Die deutschen Archive diskutieren neben der Eigenarchivierung auch die Möglichkeit, die Daten physisch in einem Rechenzentrum abzulegen (z.B. Landesarchiv Niedersachsen). Das Bundesarchiv hat bei der Wiedervereinigung zahlreiche Altdaten der DDR übernommen. Neben der Speicherung müssen die digitalen Unterlagen auch in ein zu entwickelndes Verhältnis mit den herkömmlichen papiernen Archivalien gesetzt werden, zumal auf absehbare Zeit viele Unterlagen weder rein digitaler noch ausschließlich analoger sondern hybrider Natur sein werden.

5.) Benutzung:

Archive bergen im Regelfall Unikate, die nicht ersetzt und daher nur im Lesesaal benutzt werden können. Nachdem digitale Archivalien weder den Begriff des Originals noch eine Bindung auf einen Träger kennen, können diese Archivalien auch in einem geschützten Intranet oder im Internet benutzt werden. Benutzungsmöglichkeiten über das Internet bieten derzeit die National Archives, Kew, (NDAD: http://www.ndad.nationalarchives.gov.uk/) und die NARA, Washington an (AAD: http://aad.archives.gov/aad/).

Zusammenfassend sind die deutschen Archive im Bereich der Systemeinführung sehr gut aufgestellt. In den Bereichen der Übernahme, Archivierung und Benutzung sind die angelsächsischen Archive und hier insbesondere die Nationalarchive der USA, des UK und von Australien sehr aktiv. Einen interessanten Ansatz verfolgen die staatlichen Archive der Schweiz: Sie haben 2005 auf der Grundlage einer Strategiestudie eine Koordinierungs- und Beratungsstelle (KOST) eingerichtet, die kooperative Antworten auf die digitalen Herausforderungen finden soll, http://www.vsa-aas.org/index.php?id=110&L=0.

3.3 Museen

Winfried Bergmeyer

Im Jahre 2006 gab es über 6.100 Museen und Sammlungen in Deutschland. Die Spannbreite der musealen Sammlungspolitik umfasst Werke der bildenden Kunst, historische Objekte, technische Denkmäler bis hin zu Spezialsammlungen von Unternehmen und Privatsammlern. Diese Vielfältigkeit spiegelt sich auch in den Arbeitsaufgaben der einzelnen Museen wieder. Sammeln, Bewahren, Forschen und Vermitteln als Kernbereiche der Institutionen benötigen und produzieren unterschiedlichste Informationen und dies zunehmend in digitaler Form. Nur mit digitalen Daten kann der Forderung nach schneller Verfügbarkeit und freiem Zugang zu unserem Kulturerbe in Zukunft Rechnung getragen werden. Kooperationen in Form von Projekten oder Internet-Portalen bilden dabei ein wichtiges Element der institutionsübergreifenden Erschließung von Beständen.

1. Digitale Kunst

Spätestens seit der Entwicklung der Video-Kunst ist eine Abhängigkeit der Kunstwerke von elektronischen Medien gegeben. Diese Nutzung elektronischer und digitaler Medien in der Kunst stellt die sammelnden Institutionen vor neue Herausforderungen. Hierbei geht es nicht allein um die Konservierung von Bitströmen, sondern auch von komplexen Installationen mit entsprechender Hardware. Die künstlerische Wirkung dieser Installationen wird häufig durch die spezifische Hardware zur Wiedergabe bestimmt. Die Langzeitarchivierung digitaler Kunst ist eine Herausforderung, die auf Grund ihrer Komplexität zahlreiche unterschiedliche Lösungskonzepte hervorgebracht hat. Der Ansatz, den Künstler/die Künstlerin in den Prozess der Konservierung einzubinden, ist dabei ein richtungsweisender Ansatz. In Absprache mit ihm/ihr sollte geklärt werden, wie das Verhältnis zwischen physischer Präsentationsumgebung (Hardware, Software) und inhaltlichem Konzept zu gewichten ist. Auf dieser Basis kann danach entschieden werden, welche Archivierungskonzepte gewählt werden können. Die statische Konservierung beinhaltet die Aufbewahrung (und Pflege) von Hard- und Software, also des kompletten Systems und ist die aufwändigste, technisch komplexeste und eine sicherlich nicht für alle Institutionen realisierbare Methode. Die Migration der Daten vom alten Dateiformat in ein neues, aktuelles Dateiformat oder die Emulation von Hard- und Software-Umgebungen sind alternative Konzepte zur Langzeitarchivierung. Unabhängig von der gewählten Methode ist die Forderung nach Archivierung von Infor-

mationen, die zu diesem Kunstwerk, seiner Entstehung und Rezeptionen in Beziehung stehen, für eine erfolgreiche Konservierung unerlässlich.

2. Multimediale Anwendungen

Museen sind Orte des offenen Zugangs zur kulturellen, technologischen und wissenschaftlichen Geschichte und Gegenwart. Sie vermitteln der interessierten Öffentlichkeit wissenschaftliche Informationen. In diesem Handlungsbereich erfreut sich moderne Informationstechnologie in Form von Terminalanwendungen, Internet-Auftritten und elektronischen Publikationen zunehmend größerer Beliebtheit. Die Nutzung der neuen Medien für interaktive Anwendungen ermöglicht neue Formen der Präsentation. In diesem Rahmen werden zunehmend Technologien verwendet, die sich unterschiedlicher und zum Teil kombinierter Medientypen (Audio, Video, Animationen etc.) bedienen. Hinsichtlich der Erhaltung und des langfristigen Zugriffs gibt es momentan noch wenige Konzepte und Erfahrungen. Als Bestandteil temporärer Ausstellungen werden sie häufig nach deren Ende beiseite gelegt, ohne die Möglichkeiten einer weiteren oder späteren Nutzung zu bedenken. Als Teil der Vermittlungsgeschichte oder in Form einer Nachnutzung in anderen Bereichen sollte auch, unter Beachtung von festgelegten Auswahlkriterien, hier ein Konzept zur Langzeitarchivierung bestehen. Die Komplexität und Vielfältigkeit dieser Anwendungen erfordert dabei individuelle Konzepte. Vergleichbar der Vorgehensweise bei digitaler Kunst ist besonderer Wert auf umfangreiche Dokumentation zu legen, in der die Programmierungs-Dokumentationen, Hardware-Anforderungen, Installationsvorgaben und Bedienungsanleitungen gesichert werden.

3. Sammlungsmanagement

Zu den originären Aufgaben eines Museums gehört das Sammlungsmanagement, das neben der wissenschaftlichen Inventarisation auch zahlreiche administrative Bereiche umfasst. Die digitale Inventarisation hat seit den 1990er Jahren Einzug in große und mittlere Institutionen gehalten und wird mittlerweile vermehrt von den Museumsträgern eingefordert. Sie ist integraler Bestandteil der täglichen Museumsarbeit geworden und eine wesentliche Voraussetzung für die Nutzung und Pflege der Sammlungen. Zur langfristigen Erhaltung des Wissens über die musealen Objekte ist die Erhaltung der Metadaten und ihrer Struktur notwendig. Um hier eine Langzeitverfügbarkeit zu gewährleisten sind Standards im Bereich der Ontologien, Thesauri und Vokabularien unabdingbar. Als bekanntestes Metadaten-Schema findet das der Dublin Core Metadata Initiative (http://dublincore.org) häufig Anwendung. Mit dem Datenaustauschformat

Museumdat, basierend auf dem von J. Paul Getty Trust zusammen mit ARTstor entwickelten CDWA Lite sowie dem CIDOC-CRM gibt es neue Ansätze zur Vereinheitlichung des Austauschformates komplexerer Metadaten. Die zahlreichen unterschiedlichen Vokabularien und Thesauri zur Erschließung bedürfen ebenso einer Zusammenfassung, um sammlungsübergreifendes Retrieval zu erlauben. Eine Vielzahl an Software-Herstellern bieten kleine bis große Lösungen für das Datenmanagement an. Die wichtigsten Anbieter sind mittlerweile in der Lage Schnittstellen für Metadaten nach Dublin Core anzubieten.Web-Services für Vokabularien (z.B. http://www.museumsvokabular.de) erlauben in naher Zukunft vielleicht auch hier eine Vereinheitlichung.

4. Restaurierung und Konservierung

Die Restaurierung ist in vielen Museen eine eigene Abteilung, deren Aufgabe der langfristige Erhalt der musealen Objekte ist. Die neuen Medien bieten den Restauratoren und Wissenschaftlern zahlreiche neue Möglichkeiten ihre Arbeit zu verbessern. Neben den digitalen Restaurierungsberichten bildet die Technik der virtuellen Rekonstruktion eine Methode, museale Objekte ohne Beeinträchtigung des realen Objektes zu ergänzen. Durch Nutzung virtueller Abbilder und Repräsentationen (z. B. 3D-Objekte) kann die mechanische und klimatische Belastung von empfindlichen Museumsobjekten reduziert und somit deren Erhaltung für zukünftige Untersuchungen gesichert werden. Digitale Repräsentationen sind auch als „Sicherungskopien“ für den Notfall zu verwenden. Objekte aus fragilen Materialien unterliegen oft einem nur hinauszuzögerndem Verfallsprozess, so dass hochauflösende digitale Scans hier eine konservatorische Alternative bieten. Digitalisate können natürlich nicht reale Objekte ersetzen, erlauben aber für den Fall des Verlusts eine visuelle Sicherungskopie zu erstellen, die selbstverständich nur bei entsprechender Langzeitarchivierung ihre Aufgabe erfüllen kann.
Die Komplexität und Vielschichtigkeit der in den Museen anfallenden digitalen Daten erfordern von den Institutionen ein speziell für die Sammlung definiertes Konzept für die Langzeitarchivierung. Notwendig sind individuelle Konzepte auf Basis bestehender Standards und Empfehlungen, die den personellen, finanziellen und technischen Ressourcen wie auch der jeweiligen Sammlungsstrategie gerecht werden. Dabei ist die Dokumentation der Archivierungskonzepte und ihrer Umsetzung unabdingbar.

Literatur

Staatliche Museen zu Berlin – Preußischer Kulturbesitz, Institut für Museumsforschung (Hrsg.): Statistische Gesamterhebung an den Museen der Bundesrepublik Deutschland für das Jahr 2005, Materialien aus dem Institut für Museumskunde, Heft 60, Berlin 2007

Hünnekens, Annette: Expanded Museum. Kulturelle Erinnerung und virtuelle Realitäten, Bielefeld 2002.

Depocas, Alain; Ippolito, Jon; Jones, Caitlin (Hrsg.): The Variable Media Approach - permanence through change, New York 2003

Rinehart, Richard: The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Century, http://switch.sjsu.edu/web/v6n1/article_a.htm (31.08.2007)

Witthaut, Dirk unter Mitarbeit von Zierer, Andrea; Dettmers, Arno und Rohde-Enslin, Stefan: Digitalisierung und Erhalt von Digitalisaten in deutschen Museen, *nestor*-Materialien 2, Berlin 2004

Rotheberg, Jeff: Avoiding Tecnological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, http://www.clir.org/PUBS/reports/rothenberg/contencts.html (2.9.2007)

4 Rahmenbedingungen für die Langzeitarchivierung digitaler Objekte

Stefan Strathmann

Die Langzeitarchivierung digitaler Objekte bedarf umfangreicher und verbindlicher Regelungen, die eine geordnete und dauerhafte Bereitstellung des digitalen Kulturerbes ermöglichen.

Diese Regelungen werden mit dem Begriff Policy zusammengefasst; dieser englische Begriff entspricht in diesem Zusammenhang etwa den deutschen Begriffen „Rahmenbedingungen", „Grundsätze", „Richtlinien". Bei einer Preservation Policy handelt es sich um den Plan zur Bestandserhaltung. Im Gegensatz zu einer Strategie, die festlegt, wie die Erhaltung erfolgen soll, wird von der Policy festgelegt, was und warum etwas für wie lange erhalten werden soll[1]. Die Preservation Policy ist also notwendige Grundlage für jede Preservation Strategie.

Diese Richtlinien sind nicht zeitlich befristet, sondern auf dauerhaften Bestand angelegt. D. h. sie sind, anders als beispielsweise Strategien zur Erhaltung digitaler Objekte, nicht an technischen Innovationszyklen oder politischen Veränderungen bzw. institutionellen Führungswechseln orientiert, sondern sollten langfristig Geltung haben.

1 Vgl.: Foot (2001), S. 1

Preservation Policies werden üblicherweise anhand ihres Geltungsbereiches unterschieden. Am geläufigsten sind nationale oder institutionelle Preservation Policies. Aber auch internationale Policies werden entwickelt und können maßgeblich zur Erarbeitung und Umsetzung nationaler Policies beitragen. Ein herausragendes Beispiel für eine internationale Policy ist die „Charta zur Bewahrung des digitalen Kulturerbes“[2], die am 17. Oktober 2003 auf der 32. Generalkonferenz der UNESCO verabschiedet wurde.

2 UNESCO (2003)

4.1 Nationale Preservation Policy

Stefan Strathmann

Eine nationale Preservation Policy bestimmt den Rahmen für die Bemühungen eines Staates zur Sicherung der digitalen kulturellen und wissenschaftlichen Überlieferung.
Eine solche Policy muss nicht in geschlossener Form eines Dokumentes vorliegen, vielmehr wird sie sich im Normalfall aus einer Vielzahl von Gesetzen, Bestimmungen, Vereinbarungen, Regeln etc. konstituieren.

Eine nationale Preservation Policy kann Regelungen zu sehr unterschiedlichen Fragen der digitalen Langzeitarchivierung umfassen; so finden sich typischerweise Aussagen zu verschiedenen Themenkomplexen:

- Generelles Bekenntnis, das digitale Erbe zu sichern
  Ausgangspunkt einer jeden Preservation Policy ist die verbindliche Aussage, digitale Objekte langfristig zu erhalten. Ein Staat, der den Langzeiterhalt digitaler Objekte als Aufgabe von nationaler Bedeutung erkannt hat, sollte diesem Interesse Ausdruck verleihen und so die daraus resultierenden Aktivitäten begründen und unterstützen.

- Verfügbarkeit und Zugriff
  Da die digitale Langzeitarchivierung kein Selbstzweck, sondern immer auf eine spätere Nutzung/Verfügbarkeit ausgerichtet ist, sollte dieser Bereich in einer nationalen Policy maßgeblich berücksichtigt werden. Die Rahmenbedingungen sollen eine spätere Nutzung ermöglichen.

- Rechtliche Rahmenbedingungen
  Die digitale Langzeitarchivierung ist in vielerlei Hinsicht von Rechtsfragen tangiert. Dies sollte seinen Niederschlag in allen relevanten Bereichen der Gesetzgebung finden. Hierzu gehören beispielsweise die Archivgesetze, Urheber- und Verwertungsrechte, Persönlichkeitsrechte etc.

- Finanzierung
  Eng verknüpft mit den rechtlichen Rahmenbedingungen sind auch die Fragen der Finanzierung digitaler Langzeitarchivierung. Hierzu gehört die langfristige Bereitstellung der Mittel, um die Langzeitarchivierung im gewünschten Umfang durchzuführen.

- Verantwortlichkeiten und Zuständigkeiten

Bestandteil einer nationalen Preservation Policy sind auch Festlegungen bezüglich der Verantwortlichkeiten und Zuständigkeiten. In Deutschland beispielsweise sind die Zuständigkeiten von Bund, Ländern und Gemeinden zu berücksichtigen. Vorstellbar sind auch Aussagen zur Verantwortlichkeit für bestimmte Objekttypen (Webseiten, Archivgut, Wissenschaftliche Rohdaten, Doktorarbeiten) oder fachliche Inhalte (Wissenschaftliche Literatur bestimmter Fächer).

- Auswahlkriterien
  Es sollte festgelegt sein, welche digitalen Objekte bewahrt werden sollen. Hierbei sollte das ganze Spektrum digitaler Objekte berücksichtigt werden. Da der komplette Erhalt aller digitalen Objekte kaum sinnvoll und machbar ist, sind insbesondere transparente Entscheidungs- und Auswahlkriterien von großer Wichtigkeit.

- Sicherheit
  Der Anspruch an die Sicherheit (Integrität, Authentizität, Redundanz etc.) der digitalen Bestandserhaltung sollte in einer nationalen Policy berücksichtigt werden.

In vielen Staaten finden Diskussionen zur Entwicklung nationaler Policies statt. Da zur Entwicklung einer tragfähigen nationalen Policy ein breiter gesellschaftlicher, politischer und fachlicher Konsens notwendig ist, ist die Entwicklung ein sehr langwieriger und komplizierter Prozess, der bisher nur wenig greifbare Ergebnisse aufweisen kann. Ein Beispiel für eine niedergelegte generelle nationale Preservation Policy findet sich in Australien[3]. Ein weiteres Beispiel für einen Teil einer nationalen Preservation Policy ist das „Gesetz über die Deutsche Nationalbibliothek“[4] vom 22. Juni 2006, in dem der Sammelauftrag der DNB auf Medienwerke in unkörperlicher Form (d.h. u.a. Webseiten) ausgedehnt wird. Dieses Gesetz ist selbstverständlich nicht die deutsche nationale Preservation Policy, es ist aber ein Baustein zur Definition der Rahmenbedingungen der digitalen Langzeitarchivierung in Deutschland.

In Deutschland bemüht sich insbesondere nestor um die Entwicklung einer nationalen Preservation Policy. Zu diesem Zweck wurden von nestor mehrere Veranstaltungen (mit)organisiert, eine Expertise in Auftrag gegeben[5], eine Befragung zu den Auswahlkriterien und Sammelrichtlinien durchgeführt, sowie ein

3 AMOL (1995)

4 DNBG (2006)

5 Hilf, Severiens (2006)

„Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland“[6] veröffentlicht, das sehr breit mit der Fachcommunity abgestimmt ist.

6 nestor (2006a)

4.2 Institutionelle Preservation Policy

Stefan Strathmann

Rahmenbedingungen und Grundsätze für die digitale Langzeitarchivierung müssen gemäß ihrer Dringlichkeit formuliert werden. Hierbei ist nicht nur der (inter)nationale, sondern auch der lokale und institutionsspezifische Rahmen zu berücksichtigen.
Jede mit dem Erhalt des digitalen wissenschaftlichen und kulturellen Erbe betraute Institution sollte die eigenen Grundsätze in einer institutionellen Preservation Policy festlegen. Diese Policy entspricht häufig einer Selbstverpflichtung, auch wenn weite Teile bspw. durch gesetzliche Anforderungen vorgegeben sind.
Eine solche Policy ist für die jeweiligen Institutionen dringend notwendig, um nach Innen das Bewusstsein für die Aufgaben und Belange der digitalen Langzeitarchivierung zu schaffen und nach Außen die für Vertrauenswürdigkeit notwendige Transparenz zu gewährleisten[7].

Da innerhalb einer einzelnen Institution die Abstimmungs- und Konsensfindungsprozesse häufig einfacher sind als auf nationalem Level, gibt es eine Reihe von Beispielen von institutionellen Preservation Policies[8]. Dennoch ist es bisher nicht der Regelfall, dass Gedächtnisorganisationen eine eigene Policy zum Erhalt ihrer digitalen Bestände formulieren.

Institutionelle Policies können sehr viel spezifischer an die Bedürfnisse der jeweiligen Institutionen angepasst werden, als das bei einer eher generalisierenden nationalen Policy der Fall ist. Aber auch hier ist zu bedenken, dass es sich um Leitlinien handelt, die nicht regelmäßig an das Alltagsgeschäft angepasst werden sollten, sondern dass sich vielmehr das Alltagsgeschäft an den in der Policy festgelegten Linien orientieren sollte.

Die institutionelle Preservation Policy bestimmt den Rahmen für die institutionelle Strategie zum Erhalt der digitalen Objekte. Sie sollte konkret am Zweck und Sammelauftrag der Institution ausgerichtet sein. Hierzu gehören sowohl der Sammlungsaufbau wie auch die Bedürfnisse der jeweiligen intendierten Nutzergruppen. Eine wissenschaftliche Bibliothek bspw. muss ihren Nutzern eine andere Sammlung und anderen Zugang zu dieser Sammlung zur Verfü-

7 Vgl.: nestor (2006b)

8 Vgl. bspw.: NAC (2001), OCLC (2006), PRO (2000), UKDA (2005)

gung stellen als ein Stadtarchiv oder ein Museum.

Die in den Rahmenbedingungen spezifizierten Prinzipien des Sammlungsaufbaues sollten ggf. durch Hinweise auf Kooperationen und/oder Aufgabenteilungen ergänzt werden.

Ein weiterer zentraler Bestandteil der Rahmenbedingungen für die Erhaltung digitaler Objekte innerhalb einer Institution ist die Sicherstellung der finanziellen und personellen Ressourcen für den beabsichtigten Zeitraum der Langzeitarchivierung. Eine einmalige Anschubfinanzierung ist nicht ausreichend.

Da Institutionen häufig nur eine begrenzte Zeit ihren Aufgaben nachkommen, sollte eine institutionelle Policy auch auf die Eventualitäten einer Institutionsschließung o.ä. eingehen (Fallback-Strategie, Weitergabe der archivierten Objekte an andere Institutionen).

Nutzungsszenarien sind gleichfalls wichtige Bestandteile einer institutionellen Preservation Policy. Abhängig vom Zweck der Institution sollte eine generelle Festlegung erfolgen, was wem unter welchen Bedingungen und in welcher Form zur Nutzung überlassen wird.

Fragen der Sicherheit der Daten können ebenfalls in einer institutionellen Policy geregelt werden. Dies erfolgt häufig in Form von eigens hierzu erstellten Richtlinien-Dokumenten, die Bestandteil der institutionellen Policy sind (Richtlinien zum Datenschutz, zur Netzwerksicherheit, zur Computersicherheit, zum Katastrophenschutz etc.). Auch sollte der für die Zwecke der Institution benötigte Grad an Integrität und Authentizität der digitalen Objekte festgelegt werden. In diesem Zusammenhang kann auch das Maß der akzeptablen Informationsverluste, wie sie z.B. bei der Migration entstehen können, beschrieben werden.

In einigen institutionellen Preservation Policies[9] werden sehr detailliert die Dienste der Institution festgelegt und die Strategien zur Erhaltung der digitalen Objekte spezifiziert (Emulation, Migration, Storage-Technologie etc.). Dies bedeutet, dass diese Policies relativ häufig einer Revision unterzogen und den aktuellen technischen Anforderungen und Möglichkeiten angepasst werden müssen.

9 Vgl. bspw: OCLC 2006

Literatur

AMOL (1995): National Conservation and Preservation Policy.
http://sector.amol.org.au/publications_archive/national_policies/national_preservation_strategy

DNBG (2006): Gesetz über die Deutsche Nationalbibliothek (DNBG)
http://217.160.60.235/BGBL/bgbl1f/bgbl106s1338.pdf

Foot (2001): Building Blocks for a Preservation Policy.
http://www.bl.uk/services/npo/pdf/blocks.pdf

Hilf, Severiens (2006): Zur Entwicklung eines Beschreibungsprofils für eine nationale Langzeit-Archivierungs-Strategie - ein Beitrag aus der Sicht der Wissenschaften.
http://nbn-resolving.de/urn:nbn:de:0008-20051114021

NAC (2001): National Archives of Canada: Preservation Policy
http://www.collectionscanada.ca/preservation/1304/docs/preservationpolicy_e.pdf

nestor (2006a): Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland
http://www.langzeitarchivierung.de/modules.php?op=modload&name=Downloads&file=index&req=viewdownload&cid=9

nestor (2006b): Kriterienkatalog vertrauenswürdige digitale Langzeitarchive
http://nbn-resolving.de/urn:nbn:de:0008-2006060710

OCLC (2006): OCLC Digital Archive Preservation Policy and Supporting Documentation
http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf

PRO (2000): Public Record Office: Corporate policy on electronic records
http://www.nationalarchives.gov.uk/documents/rm_corp_pol.pdf

UKDA (2005): UK Data Archive: Preservation Policy
http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf

UNESCO (2003): Charta zur Bewahrung des digitalen Kulturerbes. http://www.unesco.de/444.html (Inoffizielle deutsche Arbeitsübersetzung der UNESCO-Kommissionen Deutschlands, Luxemburgs, Österreichs und der Schweiz)

Weitere Literatur findet sich u.a. im PADI Subject Gateway (http://www.nla.gov.au/padi/), in der nestor Informationsdatenbank (http://nestor.sub.uni-goettingen.de/nestor_on/index.php) und in der ERPANET Bibliography on Digital Preservation Policies (http://www.erpanet.org/assessments/ERPANETbibliography_Policies.pdf)

4.4 Auswahlkriterien

Andrea Hänger, Karsten Huth und Heidrun Wiesenmüller

Allgemeines

Die Auswahl digitaler Objekte geschieht auf der Basis von definierten und auf die jeweilige Institution zugeschnittenen Kriterien – beispielsweise in Form von Sammelrichtlinien, Selektions- und Bewertungskriterien oder Kriterien für die Überlieferungsbildung. Im Bibliotheks- und Museumsbereich spricht man i.d.R. von Sammlungen, die aus den Sammelaktivitäten hervorgehen, im Archivbereich dagegen von Beständen, die das Resultat archivischer Bewertung darstellen. Der Begriff der Sammlung wird nur im Bereich des nicht-staatlichen Archivguts verwendet.

Bei digitalen Langzeitarchiven, die von öffentlichen Institutionen betrieben werden, sind die Auswahlkriterien i.d.R. aus dem Gesamtauftrag der Institution abzuleiten. In einigen Fällen gibt es auch gesetzliche Grundlagen – z.B. in den Archivgesetzen, die u.a. auch die formalen Zuständigkeiten staatlicher Archive regeln, oder den nationalen und regionalen Pflichtexemplargesetzen, welche Ablieferungspflichten an bestimmte Bibliotheken festlegen.

Festgelegte, dokumentierte und offen gelegte Auswahlkriterien sind in mehrfacher Hinsicht von zentraler Bedeutung für digitale Langzeitarchive: Als praktische Arbeitsanweisung für das eigene Personal unterstützen sie einen stringenten, von individuellen Vorlieben oder Abneigungen unabhängigen Aufbau der digitalen Sammlung bzw. der digitalen Bestände. Den Nutzern, aber auch den Produzenten bzw. Lieferanten der digitalen Objekte und der allgemeinen Öffentlichkeit machen sie das Profil der digitalen Sammlung bzw. der digitalen Bestände deutlich. Anhand der veröffentlichten Auswahlkriterien können beispielsweise Nutzer entscheiden, ob ein bestimmtes digitales Langzeitarchiv für ihre Zwecke die richtige Anlaufstelle ist oder nicht. Dasselbe gilt für Produzenten digitaler Objekte, soweit es keine gesetzlichen Ablieferungs- oder Anbietungspflichten gibt. Das Vorhandensein von Auswahlkriterien stellt deshalb auch einen wichtigen Aspekt von Vertrauenswürdigkeit dar.[10] Gegenüber den Trägern wird anhand der Auswahlkriterien belegt, dass die Sammelaktivitäten dem Auftrag der Institution entsprechen. Und schließlich spielen die jeweiligen

10 Das Kriterium 1.1 im 'Kriterienkatalog Vertrauenswürdige Archive' lautet: „Das digitale Langzeitarchiv hat Kriterien für die Auswahl seiner digitalen Objekte entwickelt". Vgl. nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (2006), S. 7. Zur Vertrauenswürdigkeit digitaler Langzeitarchive allgemein s.u. Kap. 8.

Auswahlkriterien auch eine wichtige Rolle beim Aufbau von Netzwerken zur verteilten, kooperativen Langzeitarchivierung (beispielsweise im nationalen Rahmen).

Zumeist stellt die Aufnahme digitaler Objekte in die Sammlung bzw. die Bestände eine zusätzliche Aufgabe dar, die zu bestehenden Sammelaktivitäten bzw. Bewertungen für konventionelle Objekte hinzukommt. Viele Institutionen besitzen deshalb bereits Auswahlkriterien im analogen Bereich, die als Ausgangspunkt für entsprechende Richtlinien im digitalen Bereich dienen können. Mit Blick auf die Besonderheiten digitaler Objekte müssen diese freilich kritisch überprüft, abgeändert und erweitert werden. Dabei sind fünf Aspekte besonders zu beachten:

- *Spezielle Objekt- und Dokumenttypen*: Während sich für viele Arten von digitalen Objekten eine Entsprechung im konventionellen Bereich finden lässt, gibt es auch spezielle digitale Objekt- und Dokumenttypen, die in den Auswahlrichtlinien zu berücksichtigen sind. Beispielsweise besitzt eine E-Dissertation im PDF-Format ein analoges Pendant in der konventionellen, gedruckten Dissertation. Eine Entsprechung für originär digitale Objekte wie Websites oder Datenbanken lässt sich hingegen nicht in gleicher Weise finden. Deshalb ist eine Orientierung an vorhandenen konventionellen Auswahlkriterien hier nur bedingt möglich (nämlich nur für die inhaltlich-fachlichen Aspekte des Objektes).
- *Technische Anforderungen*: Anders als bei konventionellen Objekten spielen technische Anforderungen (z.B. das Dateiformat und die notwendige technische Umgebung zur Darstellung der Information) für die Abläufe im digitalen Langzeitarchiv eine wichtige Rolle. Sie sind deshalb in die Überlegungen mit einzubeziehen.
- *Veränderte Arbeitsabläufe*: Digitale Objekte sind unbeständiger als ihre papierenen Gegenstücke und weniger geduldig; sollen sie dauerhaft bewahrt werden, muss bereits bei ihrer Entstehung dafür gesorgt werden. Beispielsweise müssen Bibliotheken auf die Produzenten einwirken, damit diese ihre Publikationen in langzeitgeeigneter Form erstellen; ebenso müssen Archive bei den von ihnen zu betreuenden Behörden bereits bei der Einführung elektronischer Systeme präsent sein. Sollen Informationen aus Datenbanken oder Geoinformationssystemen archiviert werden, muss sichergestellt werden, dass vorhandene Daten bei Änderung nicht einfach überschrieben werden, sondern dass so genannte Historisierungen vorgenommen werden, die einen bestimmten Stand festhal-

ten.

- *Unterschiedliche Mengengerüste*: Die Zahl und der Umfang der theoretisch auswahlfähigen digitalen Objekte liegt häufig in deutlich höheren Größenordnungen als bei entsprechenden analogen Objekten. Beispielsweise sind Netzpublikationen sehr viel leichter zu realisieren als entsprechende Printpublikationen, so dass ihre Zahl die der gedruckten Publikationen bei weitem übersteigt. Ebenso werden zum Beispiel Statistikdaten in der Papierwelt nur in aggregierter, d.h. zusammengefasster Form als Quartals- oder Jahresberichte übernommen. In digitaler Form können jedoch auch die Einzeldaten übernommen und den Nutzern in auswertbarer Form zur Verfügung gestellt werden.
- *Schwer zu bemessender Arbeitsaufwand*: Der Umgang mit konventionellen Objekten erfolgt über etablierte Kanäle und Geschäftsgänge, so dass Aufwände gut zu messen und zu bewerten sind. Der Aufwand zur Beschaffung, Erschließung, Bereitstellung und Langzeitarchivierung digitaler Objekte ist dagegen wegen fehlender Erfahrungswerte schwer abzuschätzen.

Die letzten beiden Punkte können u.U. dazu führen, dass Auswahlkriterien für digitale Objekte strenger gefasst werden müssen als für konventionelle Objekte, sofern nicht auf anderen Wegen – beispielsweise durch den Einsatz maschineller Methoden oder zusätzliches Personal – für Entlastung gesorgt werden kann. Die zusätzliche Berücksichtigung digitaler Objekte bei den Sammelaktivitäten bzw. bei der Bewertung kann außerdem Rückwirkungen auf die Auswahlkriterien für konventionelle Objekte derselben Institution haben, indem etwa die beiden Segmente in ihrer Bedeutung für die Institution neu gegeneinander austariert werden müssen.

Die zu erarbeitenden Auswahlkriterien[11] können sowohl inhaltlich-fachlicher als auch formal-technischer Natur sein. Darüber hinaus können beispielsweise auch finanzielle sowie lizenz- und urheberrechtliche Aspekte in die Auswahlkriterien mit eingehen; die folgende Liste erhebt keinen Anspruch auf Vollständigkeit.

Inhaltlich-fachliche Auswahlkriterien

Aus inhaltlich-fachlicher Sicht kommen typischerweise drei Kriterien in Betracht:

- *Verwaltungstechnische, institutionelle oder räumliche Zuständigkeit*, z.B. eines Unternehmensarchivs für die Unterlagen des Unternehmens; eines Museums für Digitalisate eigener Bestände; des Dokumentenservers einer

11 Vgl. zum Folgenden auch die Ergebnisse einer Umfrage zu den in verschiedenen Institutionen angewendeten Auswahlkriterien, die im Rahmen der ersten Phase des nestor-Projektes durchgeführt wurde: Blochmann (2005), S. 9-31.

Universität für die dort entstandenen Hochschulschriften; einer Pflichtexemplarbibliothek für die im zugeordneten geographischen Raum veröffentlichten Publikationen.
Leitfrage: Ist mein Archiv gemäß der institutionellen oder rechtlichen Vorgaben zur Übernahme des Objekts verpflichtet?

- *Inhaltliche Relevanz*, ggf. in Verbindung mit einer Qualitätsbeurteilung, z.B. thematisch in ein an einer Bibliothek gepflegtes Sondersammelgebiet fallend; zu einer Spezialsammlung an einem Museum passend; von historischem Wert für die zukünftige Forschung; von Bedeutung für die retrospektive Verwaltungskontrolle und für die Rechtssicherung der Bürger. Dazu gehört auch der Nachweis der Herkunft des Objekts aus seriöser und vertrauenswürdiger Quelle. Ggf. können weitere qualitative Kriterien angelegt werden, z.B. bei Prüfungsarbeiten die Empfehlung eines Hochschullehrers.
  Leitfragen: Ist das Objekt durch sein enthaltenes Wissen bzw. seine Ästhetik, Aussagekraft o.ä. wichtig für meine Institution? Kann das Objekt bei der Beantwortung von Fragen hilfreich sein, die an meine Institution gestellt werden? Ist das Objekt aufgrund seiner Herkunft, seiner Provenienz von bleibendem (z.B. historischem) Wert?
- *Dokumentart*, z.B. spezifische Festlegungen für Akten, Seminararbeiten, Geschäftsberichte, Datenbanken, Websites etc.
  Leitfragen: Besitzt mein Archiv schon Bestände der gleichen Dokumentart? Verfüge ich über das nötige Fachwissen und die nötigen Arbeitsmittel zur Erschließung und Verzeichnung der Dokumentart?

Formal-technische Auswahlkriterien

Aus formal-technischer Sicht steht auf der obersten Ebene das folgende Kriterium:

- *Lesbarkeit des Objekts im Archiv*, z.B. die Prüfung, ob ein Objekt mit den verfügbaren technischen Mitteln (Hardware/Software) des Langzeitarchivs dargestellt werden kann. Darstellen heißt, dass die vom Objekt transportierte Information vom menschlichen Auge erkannt, gelesen und interpretiert werden kann.
  Leitfrage: Verfügt mein Archiv über die nötigen Kenntnisse, Geräte und Software, um das Objekt den Nutzern authentisch präsentieren zu können?

Aus diesem obersten formal-technischen Zielkriterium lassen sich weitere Unterkriterien ableiten:

- *Vorhandensein der notwendigen Hardware*, z.B. die Feststellung, ob ein einzelner Rechner oder ein ganzes Netzwerk benötigt wird; ob die Nutzung

des Objekts an ein ganz spezielles Gerät gebunden ist usw. Außerdem muss geprüft werden, ob das Objekt mit den vorhandenen Geräten gespeichert und gelagert werden kann.
Leitfragen: Verfügt mein Archiv über ein Gerät, mit dem ich das Objekt in authentischer Form darstellen und nutzen kann? Verfügt mein Archiv über Geräte, die das Objekt in geeigneter Form speichern können?

- *Vorhandensein der notwendigen Software*, z.B. die Feststellung, ob die Nutzung eines Objekts von einem bestimmten Betriebssystem, einem bestimmten Anzeigeprogramm oder sonstigen Einstellungen abhängig ist. Außerdem muss das Archiv über Software verfügen, die das Speichern und Auffinden des Objektes steuert und unterstützt.
  Leitfragen: Verfügt mein Archiv über alle Programme, mit denen ich das Objekt in authentischer Form darstellen und nutzen kann? Verfügt mein Archiv über Programme, die das Objekt in geeigneter Form speichern und wiederfinden können?
- *Vorliegen in geeigneten Formaten*, bevorzugt solchen, die normiert und standardisiert sind, und deren technische Spezifikationen veröffentlicht sind. Dateiformate sollten nicht von einem einzigen bestimmten Programm abhängig, sondern idealerweise weltweit verbreitet sein und von vielen genutzt werden. Je weniger Formate in einem Archiv zulässig sind, desto leichter kann auch das Vorhandensein der notwendigen Hard- und Software geprüft werden.
  Leitfragen: Hat mein Archiv schon Objekte dieses Formats im Bestand? Sind die notwendigen Mittel und Kenntnisse zur Nutzung und Speicherung des Formats offen zugänglich und leicht verfügbar?
- *Vorhandensein geeigneten Personals*, z.B die Feststellung, ob die Mitarbeiterinnen und Mitarbeiter über das technische Fachwissen verfügen, das zur Nutzung und Speicherung des Objekts notwendig ist.
  Leitfragen: Habe ich Personal, dem ich aus technischer Sicht die Verantwortung für das Objekt anvertrauen kann? Verfüge ich über die Mittel, um Personal mit den entsprechenden Kenntnissen einzustellen oder um Dienstleister mit der Aufgabe zu betrauen?

Auswahlkriterien für Netzpublikationen

Eine für Bibliotheken besonders wichtige Gattung digitaler Objekte sind die so genannten *Netzpublikationen*, auch als „Medienwerke in unkörperlicher Form" bezeichnet und als „Darstellungen in öffentlichen Netzen"[12] definiert. Auch für diese gelten die oben dargestellten allgemeinen Auswahlkriterien, doch sollen im Folgenden noch einige spezielle Hinweise aus bibliothekarischer Sicht gegeben werden[13]. Dabei ist es nützlich, die Vielfalt der Netzpublikationen in zwei Basistypen zu unterteilen: In die Netzpublikationen mit Entsprechung in der Printwelt einerseits und die sog. Web-spezifischen Netzpublikationen andererseits.[14]

Bei den *Netzpublikationen mit Entsprechung in der Printwelt* lassen sich wiederum zwei Typen unterscheiden:

- *Druckbildähnliche Netzpublikationen*, welche ein weitgehend genaues elektronisches Abbild einer gedruckten Publikation darstellen, d.h. 'look and feel' des gedruckten Vorbilds möglichst exakt nachahmen wollen und diesem bis hin zum äußeren Erscheinungsbild entsprechen (z.B. Titelblatt, festes Layout mit definierten Schriftarten und -größen, feste Zeilen- und Seitenumbrüche etc.).
- *Netzpublikationen mit verwandtem Publikationstyp in der Printwelt*, welche zwar keine Druckbildähnlichkeit aufweisen, jedoch einem aus der Printwelt bekannten Publikationstyp zugeordnet werden können, z.B. ein Lexikon im HTML-Format.

Bei der Erarbeitung von Auswahlkriterien für diese beiden Typen ist i.d.R. eine Orientierung an bereits vorhandenen Sammelrichtlinien für konventionelle Materialien möglich. Besondere Beachtung verdient dabei der durchaus nicht seltene Fall, dass zur jeweiligen Netzpublikation eine gedruckte Parallelausgabe vorliegt. Unter Abwägung des zusätzlichen Aufwandes einerseits und des möglichen Mehrwerts des digitalen Objekts andererseits ist festzulegen, ob in einem solchen Fall nur die konventionelle oder nur die digitale Version in das Archiv aufgenommen wird, oder ob beide Versionen gesammelt werden.

Zu den *Web-spezifischen Netzpublikationen* zählen beispielsweise Websites oder Blogs. Sie können keinem aus der Printwelt bekannten Publikationstyp zugeordnet werden, so dass eine Orientierung an bestehenden Sammelrichtlinien

12 Gesetz über die Deutsche Nationalbibliothek (2006), § 3, Abs. 3.

13 Auf andere Arten von Gedächtnisorganisationen ist die folgende Darstellung nicht zwingend übertragbar.

14 Für die folgenden Ausführungen vgl. Wiesenmüller et al. (2004), S. 1423-1437. Unbenommen bleibt, dass die im Folgenden genannten Typen von Netzpublikationen auch in Offline-Versionen vorkommen können.

nur sehr bedingt möglich ist. Für diese Publikationstypen müssen daher neue Auswahlkriterien entwickelt werden.[15]

Der Umgang mit *Websites* wird dadurch erschwert, dass unterhalb der Website-Ebene häufig weitere Netzpublikationen - mit oder ohne Entsprechung in der Printwelt - liegen, die getrennt gesammelt, erschlossen und bereitgestellt werden können (z.B. ein Mitteilungsblatt auf der Website einer Institution). In den Auswahlkriterien muss also auch festgelegt sein, unter welchen Umständen (nur) die Website als Ganzes gesammelt wird, oder zusätzlich bzw. stattdessen auch darin integrierte Netzpublikationen in das Archiv aufgenommen werden sollen. Bei Websites, die immer wieder ergänzt, aktualisiert oder geändert werden und deshalb in Zeitschnitten zu sammeln sind, muss jeweils auch das Speicherintervall festgelegt werden.

Bei der Erarbeitung von Auswahlkriterien für Websites sollte unterschieden werden zwischen solchen, welche Personen oder Körperschaften (inkl. Gebietskörperschaften, Ausstellungen, Messen etc.) repräsentieren, und solchen, die sich einem bestimmten Thema widmen – wobei freilich auch Mischformen möglich sind.

Bei *repräsentierenden Websites* setzen die Auswahlkriterien in erster Linie beim Urheber an: Ist die repräsentierte Person oder Körperschaft für mein Archiv relevant? Welche Arten von Personen und Körperschaften sollen schwerpunktmäßig gesammelt, welche ausgeschlossen werden? Ein zusätzliches Kriterium können die auf der Website gebotenen Informationen sein, was sich am besten am Vorhandensein und an der Gewichtung typischer Elemente festmachen lässt: Beispielsweise könnten Websites, die umfangreiche Informationen zur repräsentierten Person oder Körperschaft, einen redaktionellen Teil und/oder integrierte Netzpublikationen bieten, mit höherer Priorität gesammelt werden als solche, die im wesentlichen nur Service- und Shop-Angebote beinhalten.

Bei *thematischen Websites* kommt neben der inhaltlichen Relevanz auch die Qualität als Auswahlkriterium in Frage. Zwar kann i.d.R. keine Prüfung auf Richtigkeit oder Vollständigkeit der gebotenen Information geleistet werden, doch

15 Auch Online-Datenbanken sind am ehesten den Web-spezifischen Netzpublikationen zuzuordnen, weil es in der Printwelt keinen Publikationstyp gibt, der in Funktionalität und Zugriffsmöglichkeiten mit ihnen vergleichbar ist. Ein grundsätzlicher Unterschied zu einem gedruckten Medium ist z.B., dass dessen gesamter Inhalt sequentiell gelesen werden kann, während bei einer Datenbank gemäß der jeweiligen Abfrage nur eine Teilmenge des Inhalts in lesbarer Form generiert wird. Was jedoch den in Online-Datenbanken präsentierten *Inhalt* angeht, so kann es natürlich durchaus Entsprechungen zu Produkten aus der Printwelt geben (z.B. sind in vielen Fällen gedruckte Bibliographien durch bibliographische Datenbanken abgelöst worden).

können als Auswahlkriterien u.a. der Umfang, die Professionalität der Darbietung und die Pflege der Website herangezogen werden, außerdem natürlich der Urheber (z.B. Forschungsinstitut vs. Privatperson).
Detaillierte Sammelrichtlinien für Netzpublikationen, die als Anregung dienen können, sind beispielsweise im Rahmen des PANDORA-Projekts von der Australischen Nationalbibliothek erarbeitet und veröffentlicht worden.[16]

Quellenangaben

Blochmann, Andrea (2005): Langzeitarchivierung digitaler Ressourcen in Deutschland: Sammelaktivitäten und Auswahlkriterien (nestor – Kompetenznetzwerk Langzeitarchivierung, AP 8.2). Version 1.0. Frankfurt am Main: nestor.

http://www.langzeitarchivierung.de/downloads/nestor_ap82.pdf (08.10.2007).

Gesetz über die Deutsche Nationalbibliothek (2006): vom 22. Juni 2006.

http://www.d-nb.de/wir/pdf/dnbg.pdf (08.10.2007).

National Library of Autralia (2005): Online Australian publications: selection guidelines for archiving and preservation by the National Library of Australia. Rev. August 2005. Canberra: National Library of Australia.

http://pandora.nla.gov.au/selectionguidelines.html (14.10.2007).

nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2006): Kriterienkatalog vertrauenswürdige digitale Langzeiarchive. Version 1 (Entwurf zur öffentlichen Kommentierung). (nestor-Materialien 8). Frankfurt am Main: nestor.

http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf (08.10.2007). urn:nbn:de:0008-2006060710.

Arbeitskreis Archivische Bewertung im VdA – Verband deutscher Archivarinnen und Archivare (Hrsg.) (2004): Positionen des Arbeitskreises Archivische Bewertung im VdA – Verband deutscher Archivarinnen und Archivare zur archivischen Überlieferungsbildung: vom 15. Oktober 2004.

http://www.vda.archiv.net/texte/ak_bew_positionen2004.doc (12.10.2007)

Wiesenmüller, Heidrun et al. (2004): Auswahlkriterien für das Sammeln von Netzpublikationen im Rahmen des elektronischen Pflichtexemplars: Empfehlungen der Arbeitsgemeinschaft der Regionalbibliotheken. In: Bibliotheksdienst 11. 2004 (Jg. 38), S. 1423-1444.

http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/digitale-bib1104.pdf (08.10.2007).

16 Vgl. National Library of Australia (2005).

5 Geschäftsmodelle

5.1 Kosten

Thomas Wollschläger

In diesem Kapitel werden Kostenfaktoren benannt, die für den Betrieb eines digitalen Langzeitarchivs von Bedeutung sind. Des Weiteren werden Ansätze vorgestellt, wie die individuellen Kosten der Langzeitarchivierung (LZA) in einer Institution ermittelt werden können.

5.1.1 Kostenfaktoren bei Einrichtung und Unterhaltung eines Langzeitarchivs

Abhängig vom konkreten Langzeitarchivierungskonzept der jeweiligen Einrichtung können folgende Kostenfaktoren zu berücksichtigen sein:

Initiale Kosten

- Informationsbeschaffung über LZA-Systeme
- Erhebung von Bestand, Zugang und gewünschten Zugriffsoptionen für digitale Materialien im eigenen Haus
- Erhebung von vorhandenen Personal- und Technikressourcen im eigenen Haus
- Projektplanung, ggf. Consulting, Ausschreibung(en)

Beschaffungskosten

- Hardware: Speichersysteme und sämtliche infrastrukturellen Einrichtungskosten (Serveranbindungen, Datenleitungen, Mitarbeiterrechner usw.)
- Ggf. Lizenz(en) für Software-Systeme oder Beitrittskosten zu Konsortien
- Weitere Aufwendungen: z.B. Anpassungsentwicklungen von Open Source Software-Produkten, Entwicklung/Anpassung von Schnittstellen, Erstellung von physischen und digitalen Schutzmaßnahmen (auch solche aus rechtlichen Gründen)
- Ggf. Einstellung neuer Mitarbeiter und/oder Schulung vorhandener Mitarbeiter

Betriebskosten

- Dateningest des bisher vorhandenen Materials
- Dateningest des neu eingehenden Materials
- Laufende Storage-Kosten
- Sonstige Dauerbetriebskosten: z.B. Strom, Datenleitungskosten, sämtliche Sicherheitsmaßnahmen, Backups, regelmäßige Wartung(en) und Tests, Software-Upgrades
- Zukauf von weiteren Speichereinheiten
- Hard- und Software-Komplettersatz in Intervallen
- Ggf. laufende Lizenzkosten und/oder laufende Beitragszahlungen bei Konsortien

Die konkreten Kosten sind dabei jeweils abhängig von

- der Zahl und Komplexität der Workflows bei einer Institution
- der Menge, Heterogenität und Komplexität der zu archivierenden Objekte und ihrer Metadaten
- den gewünschten Zugriffsmöglichkeiten und Schnittstellen sowie ggf.
- den Anforderungen Dritter an die archivierende Institutionen bzw. Verpflichtungen der Institution gegenüber Dritten

5.1.2 Die Ermittlung von Kosten für die Langzeitarchivierung

Die tatsächliche Ermittlung der Kosten, die auf eine Einrichtung für die Langzeitarchivierung ihrer digitalen Dokumente zukommen, gestaltet sich in der Praxis noch relativ schwierig. Viele LZA-Unternehmungen befinden sich derzeit (2007) noch im Projektstatus oder haben gerade mit dem produktiven Betrieb begonnen. Daher liegen kaum Erfahrungswerte vor, wie sich insbesondere der

laufende Betrieb eines solchen Archivs kostenmäßig erfassen lässt. Außerdem befinden sich nach wie vor die zunehmende Menge und Varianz insbesondere der Internet-Publikationen in einem Wettlauf mit den technischen Möglichkeiten, die von Gedächtnisorganisationen zur Einsammlung und Archivierung eingesetzt werden können.

Einen begrenzten Anhaltspunkt können die angesprochenen Unternehmungen zumindest in der Hinsicht liefern, was die Ersteinrichtung eines digitalen LZA betrifft. Das BMBF und die DFG haben eine ganze Reihe von solchen Projekten gefördert, und verschiedene Institutionen haben Projekte aus eigenen Mitteln finanziert[1]. Das bisher am umfangreichsten geförderte LZA-Vorhaben in Deutschland war das Projekt kopal mit einem Fördervolumen von 4,2 Mio. Euro[2]. Diese Kosten schließen die vollständige Entwicklung eines Archivsystems einschließlich Objektmodell, Aufbau von Hard- und Softwareumgebungen in mehreren Einrichtungen und mehrjährige Forschungsarbeiten ein. Zum Projektende hat kopal allerdings in einem Servicemodell konkrete Kosten für den Erwerb eines vollständigen Archivs zum Eigenbetrieb vorgelegt. Wenn das kopal-Archivsystem unter Zukauf von Beratung und ggf. Entwicklung eigenständig betrieben wird, soll ein Erstaufwand für Hard- und Software eines Systems mittlerer Größe von ca. 750.000 € anfallen. Hiervon entfielen 40% auf Softwarelizenzen und 60% auf Systembereitstellung und –betrieb[3]. Wiewohl solche Angaben nur exemplarisch sein können, kann dennoch davon ausgegangen werden, dass die Kosten für die Ersteinrichtung eines LZA-Systems in einer Einrichtung einen gewissen Schwellenwert nicht unterschreiten werden.

Die Zahl der Ansätze, die bisher versucht haben, Modelle für die Betriebskostenermittlung digitaler LZA zu entwickeln, ist begrenzt. Nennenswert ist hierbei der Ansatz des LIFE-Projekts aus Großbritannien. „The LIFE Project“ war ein einjähriges Projekt (2005/2006) der British Library (BL) in Zusammenarbeit mit dem University College London (UCL) mit dem Hauptziel, ein Kostenmanagement für die Langzeiterhaltung elektronischer Ressourcen zu entwickeln. Es wurde eine Formel zur Ermittlung der Archivierungskosten entwickelt. Manche Fragen mussten noch offen blieben, so war es z.B. bislang nicht adäquat möglich, im Rahmen des Projektes die Kosten der Langzeiterhaltung von gedruckten und elektronischen Veröffentlichungen zu vergleichen.

1 Siehe dazu die Projektübersicht in der nestor-Informationsdatenbank: <http://www.langzeitarchivierung.de/modules.php?op=modload&name=PagEd&file=index&page_id=16>

2 Vgl. Wollschläger (2007), S. 247.

3 Siehe kopal (2007), S. 2.

Die Formel lautet: $\mathbf{L_T=A_q+I_T+M_T+Ac_T+S_T+P_T}$. Dabei stehen die Werte für folgende Parameter[4]:

- L = complete lifecycle cost over time 0 to T.
- Aq = Acquisition
- I = Ingest
- M = Metadata
- Ac = Access
- S = Storage
- P = Preservation

Jeder der Parameter kann weiter in praktische Kategorien und Elemente aufgeteilt werden. Alle Schritte können entweder, wenn der Prozess direkt kalkulierbar ist, als Kostenfaktor berechnet werden oder, wenn nötig, jeweils auch noch in beliebig viele Unterpunkte untergliedert werden. So kann die Berechnung für die jeweilige Institution individuell angepasst werden. Innerhalb des LIFE-Projekts wurden zum einen beispielhafte Berechnungen der LZA-Kosten des Projektmaterials vorgenommen und dabei Kosten für „the first year of a digital asset's existence" und „the tenth year of the same digital assets' existence" vergleichbar ermittelt[5] und exemplarisch Kosten pro Speichermenge. Zum anderen hat das Projekt die entwickelten Formelwerke zur Verfügung gestellt, so dass interessierte Institutionen selbst Berechnungen anhand der Individualparameter vornehmen können.

Eine bedeutende Frage für die Festlegung der Archivierungsstrategie – nämlich für das eigentliche „Preservation Planning", die Erhaltungsmaßnahmen über die Lebenszeit eines digitalen Objekts – einer Institution ist, ob auf Dauer Migrationen oder Emulationen kostengünstiger sind. Hierzu sind noch keine abschließenden Aussagen möglich. Generell verbreitet ist die Auffassung, dass Migration der kostengünstigere Weg sei. Innerhalb von LIFE wurden dazu Ansätze formuliert, die jedoch hauptsächlich sehr exemplarische Migrationen behandeln und noch nicht repräsentativ sind[6]. Andere Studien kommen dagegen zu dem Schluss, dass Emulationen auf längere Sicht kostengünstiger seien:

> *While migration applies to all objects in the collection repetitively, emulation applies to the entire collection as a whole. This makes emulation most cost-effective in cases of large collections, despite the relatively high initial costs for developing an emulation device. When considering the fact that only small fragments of digital archives need to*

4 Vgl. McLeod/Wheatley/Ayris (2006), S. 6.

5 Vgl. Ebenda, S. 3.

6 Vgl. Ebenda, S. 10.

be rendered in the long run, it may turn out that from a financial perspective emulation techniques will be more appropriate for maintaining larger archives[7].

Da die bestehenden Langzeitarchive gerade erst dabei sind, die ersten „echten" Maßnahmen von Preservation Planning umzusetzen, wird hier auf Erfahrungswerte zu warten sein, die entsprechende Ergebnisse unterstützen können.

5.1.3 Konsequenzen für die Gedächtnisorganisationen

Angesichts der zu erwartenden nicht unerheblichen Kosten für die *Ersteinrichtung* eines LZA-Systems dürften kleinere Einrichtungen nicht umhin kommen, zwecks Einrichtung eines solchen Systems mit anderen Institutionen zu kooperieren bzw. sich einem bestehenden System anzuschließen und/oder sich den Zugang dazu über Lizenzen zu sichern. Selbst größere Institutionen werden für die Einrichtung eines LZA-Systems oft kooperative Formen wählen, um hohe Ersteinrichtungskosten, die sich sonst nicht auf mehrere Schultern verteilen lassen, zu vermeiden. Ebenso könnte angesichts der noch bestehenden Unsicherheit, wie sich künftig die Kosten für den Dauerbetrieb des Langzeitarchivs und das Preservation Planning entwickeln werden, die Entscheidung zugunsten der Variante ausfallen, sich in bestehende Systeme einzukaufen oder über kostenpflichtige Lizenzen Teilnehmer an einem kommerziell ausgerichteten System zu werden. Letzteres macht in der Regel Zugeständnisse an die gewünschte Preservation Policy notwendig, so dass eine Gedächtnisorganisation abwägen muss, welche Kosten – Lizenzen für ein kommerzielles System oder eigene Entwicklungskosten, z.B. für die Anpassung von Open Source Software – die jeweils lohnendere Investition ist.

Die Teilnahme an kooperativen Formen der Langzeitarchivierung ist unter Kostenaspekten in jedem Fall empfehlenswert. Hierbei können Institutionen über z.B. gemeinsame Speichernutzung bzw. gegenseitiges Backup, gegenseitige Nutzung von Entwicklungsergebnissen, gemeinsame Adressierung übergreifender Herausforderungen oder kooperative Verwaltung von Open Source Software Synergien schaffen und erhebliche Ressourceneinsparungen ermöglichen.

7 Zitiert nach Oltmans/Kol (2005), #5 – Conclusion.

Literatur

Kopal (2007): kopal: Ein Service für die Langzeitarchivierung digitaler Informationen. Projekt kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen), 2007 (s. < http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf>)

McLeod, Rory / Wheatley, Paul / Ayris, Paul (2006): Lifecycle Information for E-literature : A summary from the LIFE project ; Report Produced for the LIFE conference 20 April 2006. LIFE Project, London (via <http://www.ucl.ac.uk/ls/lifeproject/> or directly under <http://eprints.ucl.ac.uk/archive/00001855/01/LifeProjSummary.pdf>)

Oltmans, Erik / Kol, Nanda (2005): A Comparison Between Migration and Emulation in Terms of Costs. In: RLG DigiNews, Volume 9, Number 2, 15.04.2005 (<http://www.rlg.org/en/page.php?Page_ID=20571>).

Wollschläger, Thomas (2007): kopal – ein digitales Archiv zur dauerhaften Erhaltung unserer kul-turellen Überlieferung. In: Geschichte im Netz : Praxis, Chancen, Visionen ; Beiträge der Tagung .hist2006, Berlin: Clio-online und Humboldt-Universität zu Berlin, 2007, S. 244 – 257 (Historisches Forum 10 (2007), Teilband I).

Siehe außerdem die Einträge in der nestor-Informationsdatenbank zum Thema „Kosten" unter <http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?show=8>.

5.2 Service- und Lizenzmodelle

Thomas Wollschläger

In den wenigsten Fällen werden Langzeitarchivierungssysteme von einer einzigen Institution produziert und genutzt. Schon bei einer zusätzlichen Nutzer- oder Kundeninstitution für das hergestellte und/oder betriebene Archivsystem müssen Lizenz- oder Geschäftsmodelle aufgestellt sowie Servicemodelle für zu leistende Langzeitarchivierungs-Dienstleistungen definiert werden.

5.2.1 Lizenzmodelle

Lizenzkosten fallen in der Regel für die Nutzung kommerzieller Softwareprodukte an. Dabei gibt es unterschiedliche Möglichkeiten. Zum einen können solche Produkte lizenziert und eigenständig in der eigenen Institution eingesetzt werden. Dabei ist die Hersteller- oder Vetriebsfirma neben den (einmalig oder regelmäßig) zu zahlenden Lizenzgebühren zumeist durch Support- und Updateverträge mit der Nutzerinstitution verbunden. Beispiele hierfür sind etwa das System *Digitool* der Firma Exlibris[8] oder das *DIAS*-System von IBM[9].

Zum anderen besteht aber bei einigen Produkten aber auch die Möglichkeit, dass eine Betreiberinstitution (die nicht identisch mit dem Hersteller oder Systemvertreiber sein muss) das Archivsystem hostet und eine Nutzung für Dritte anbietet. Hierbei werden Lizenzkosten meist vom Betreiber auf die Kunden umgelegt oder fließen in die Nutzungskosten für die Archivierung ein. Ein Beispiel hierfür ist das insbesondere auf die Archivierung von e-Journals ausgerichtete System *Portico*. Hierbei erfolgt eine zentrale, an geografisch auseinander liegenden Orten replizierte Archivierung. Die Kosten von Portico richten sich für eine Bibliothek nach dem verfügbaren Erwerbungsetat. Der jährliche Beitrag für die Nutzung des Systems kann daher je nach dessen Höhe zwischen 1% des Erwerbungsetats und maximal 24.000 US-$ liegen[10].

Neben den kommerziellen Produkten gibt es eine Reihe von Open Source – Lösungen im Bereich der Archivierungssysteme. Durch die Nutzung von

8 Siehe <http://www.exlibrisgroup.com/digitool.htm>

9 Siehe <http://www-05.ibm.com/nl/dias/>

10 Vgl. <http://www.portico.org/libraries/aas_payment.html>

Open Source – Lizenzen[11] fallen oft keine Lizenzgebühren bzw. –kosten für die Nutzerinstitutionen an, sondern zumeist nur Aufwands- und Materialkosten. Zudem sind Archivinstitutionen, die eine Open Source – Software oder ein Open Source – Netzwerk nutzen, dahingehend gefordert, durch eigene Entwicklungsbeiträge das Produkt selbst mit weiterzuentwickeln[12]. Beispiele für verbreitete Open Source – Lösungen sind das System *DSpace*[13] und die *LOCKSS*- bzw. *CLOCKSS*-Initiative[14]. Die LOCKSS-Technologie will die langfristige Sicherung des archivierten Materials dadurch sicherstellen, dass jedes Archivobjekt mit Hilfe des Peer-to-Peer-Prinzips bei allen Mitgliedern gleichzeitig gespeichert wird. Jedes Mitglied stellt einen einfachen Rechner exklusiv zur Verfügung, der im Netzwerk mit den anderen Mitgliedern verbunden ist und auf dem die LOCKSS-Software läuft.

Neben der Nutzung reiner kommerzieller Lösungen und reiner Open Source – Lösungen gibt es auch Mischformen. Dabei kann es von Vorteil sein, nur für Teile des eigenen LZA-Systems auf kommerzielle Produkte zurückzugreifen, wenn sich dadurch beispielsweise die Höhe der anfallenden Lizenzkosten begrenzen lässt. Andererseits erwirbt man mit vielen Lizenzen zumeist auch Supportansprüche, die etwa bei geringeren eigenen Entwicklungskapazitäten willkommen sein können. Ein Beispiel für eine solche LZA-Lösung ist das *kopal*-System. Hierbei wird das lizenz- und kostenpflichtige (modifizierte) Kernsystem DIAS verwendet, während für den Ingest und das Retrieval die kostenfreie Open Source – Software *koLibRI* zur Verfügung gestellt wird[15].

Eine Institution muss somit abwägen, welches Lizenzmodell für sie am vorteilhaftesten ist. Kommerzielle Lizenzen setzen den Verwendungs- und Verbreitungsmöglichkeiten der Archivsysteme oft enge Grenzen. Open Source – Lizenzen bieten hier in der Regel breitere Möglichkeiten, verbieten aber ggf. die Exklusivität bestimmter Funktionalitäten für einzelne Institutionen. Hat sie ausreichende Entwicklungskapazitäten und Hard- bzw. Softwareausstattung, kann die Nutzung von Open Source Lösungen ein guter und gangbarer Weg sein. Dies gilt beispielsweise auch, wenn sich die Institution als Vorreiter für

11 Siehe hierzu v.a. < http://www.opensource.org/licenses>

12 Vgl. hierzu insbesondere das Kapitel „Kostenrelevante Eigenschaften einer ungewöhnlichen

Organisationsform“, in: Lutterbeck/Bärwolff/Gehring, S. 185 – 194.

13 Siehe <http://www.dspace.org/>

14 Siehe <http://www.lockss.org/>

15 Siehe <http://kopal.langzeitarchivierung.de/index_koLibRI.php.de>.

leicht nachnutzbare Entwicklungen sieht oder im Verbund mit anderen Einrichtungen leicht konfigurierbare Lösungen erarbeiten will. Hat sie jedoch nur geringe Enwicklungsressourcen und decken die kommerziellen Lizenzen alle benötigten Services ab, so kann trotz ggf. höherer Lizenzkosten die Wahl kommerzieller Produkte bzw. von standardisierten Services seitens LZA-Dienstleistern angeraten sein.

5.2.2 Servicemodelle

Wie bereits dargestellt, bestehen die entscheidenden Faktoren für die Entscheidung einer Institution für bestimmte Lizenz- und Geschäftsmodelle in den von ihr benötigten Services zur Langzeitarchivierung[16]. Entscheidungskriterien für die Wahl der Einrichtung und/oder Nutzung bestimmter LZA-Services können sein:

Auftrag und Selbstverständnis

- Liegt ein (z.B. gesetzlicher) Auftrag vor, dass die Institution digitale Dokumente eines bestimmten Portfolios sammeln und (selbst) langzeitarchivieren muss?
- Gilt dieser Auftrag auch für Materialien Dritter (z.B. durch Pflichtexemplarregelung)?
- Hat die Institution den Anspruch oder das Selbstverständnis, LZA-Services selbst anbieten oder garantieren zu wollen?
- Liegt eine rechtliche Einschränkung vor, Materialien zwecks LZA Dritten zu übergeben?

Ausstattung und Ressourcen

- Hat die Institution die benötigte Hardware- und/oder Softwareausstattung bzw. kann sie bereitstellen, um LZA betreiben zu können?
- Tritt die Institution bereits als Datendienstleister auf oder ist sie selbst von Datendienstleistern (z.B. einem Rechenzentrum) abhängig?
- Stehen genügend personelle Ressourcen für den Betrieb, den Support (für externe Nutzer) und für nötige Entwicklungsarbeiten zur Verfügung?
- Lassen die Lizenzen des genutzten Archivsystems / der Archivsoftware eine Anbindung Dritter an die eigene Institution zwecks LZA zu?

16 Selbstverständlich spielen auch die technischen Möglichkeiten des eingesetzten Archivsystems selbst eine wesentliche Rolle. Einen Kriterienkatalog zur technischen Evaluierung von Archivsystemen bietet z.B. das Kapitel *Software Systems for Archiving* bei Borghoff, S. 221 – 238.

Je nachdem, wie diese Fragen beantwortet werden, stehen für die Wahl des Servicemodells potentiell viele Varianten zur Verfügung. Diese drehen sich im Wesentlichen um die folgenden Konstellationen:

- Die Institution stellt einen LZA-Service (nur) für digitale Dokumente aus eigenem Besitz bereit.
- Die Institution stellt diesen LZA-Service auch für Dritte zur Verfügung.
- Die Institution stellt selbst keinen LZA-Service bereit, sondern nutzt die Services eines Dritten für die Archivierung der eigenen Daten.

Dabei ist jeweils zusätzlich und unabhängig von der Frage, welche den *Service* an sich anbietet, relevant, ob die Daten bzw. respektive die Hardware-/Storage-Umgebung von der Service-Institution selbst oder von Dritten gehostet wird. Beispielsweise kann eine Institution verpflichtet sein, selbst einen LZA-Service anzubieten. Dennoch mag der Umfang des jährlich anfallenden Materials den aufwändigen Aufbau einer solchen Hardware-/Storage-Umgebung sowie entsprechender Betriebskompetenzen nicht rechtfertigen. Hier könnte die Institution entschieden, zwar einen LZA-Service aufzubauen – und ggf. sogar Dritten über ein entsprechendes Geschäftsmodell anzubieten –, das Datenhosting jedoch an einen geeigneten Dienstleister abzugeben. Ein Beispiel für ein solches Servicekonzept ist das *kopal*-Projekt. Die Hauptmandanten betreiben zwar gemeinschaftlich das Archivsystem kopal und stellen ihre Dienstleistungen (zumeist kleineren) Nutzerinstitutionen zur Verfügung, die eigentliche Datenhaltung wird jedoch bei einem Rechenzentrum betrieben, wo die gemeinschaftlich genutzte Hardware zentral gehostet und per Fernzugriff genutzt werden[17].

Zu den einzelnen Dienstleistungen, die im Rahmen eines LZA-Service-Modells von einer Institution angeboten werden können, gehören beispielsweise folgende:

- Der Betrieb des LZA-Systems und Annahme von Archivmaterial
- Durchführung von Erhaltungsmaßnahmen (von Bitstream-Preservation bis zur Migration von Material)
- Zurverfügungstellung von Datenkopien bei Datenverlusten seitens der Ablieferinstitution
- Installation des Systems bzw. von Zugangskomponenten für Remote Access vor Ort

17 Siehe kopal (2007), S. 1-2.

- Beratungsleistungen, z.B. zum Geschäftsmodell, zum Einsatz der Archivsoftware, zur Speicherverwaltung etc.
- Support und Schulungen
- Weiterentwicklung des Archivsystems bzw. von gewünschten Komponenten

Handelt es sich bei dem Dienstleister, der von einer Archivinstitution in Anspruch genommen wird, um einen reinen Datenhost, könnten folgende Dienstleistungen relevant werden:

- Hardwarehosting und –betreuung
- Hosting und Betreuung von Standardsoftware
- Sichere Datenhaltung (z.B. durch Mehrfachbackups)
- Zurverfügungstellung von Datenkopien bei Datenverlusten seitens der Ablieferinstitution
- Notfall- und Katastrophenmanagement
- Beratungsleistungen, z.B. zur Speicherverwaltung

Jede Institution muss die eigenen Möglichkeiten bezüglich des Angebots von LZA-Services sorgfältig evaluieren. Hat sie einmal damit begonnen, insbesondere für Dritte solches Services anzubieten, werden dadurch Verpflichtungen eingegangen, die durch künftige technische Entwicklungen ggf. nur erschwert eingehalten werden können. Daher kann es ratsam sein, LZA-Services koordiniert oder kooperativ mit anderen Einrichtungen anzubieten bzw. zu nutzen. Lassen sich die Dienstleistungen von externen Anbietern nutzen und ist dies auch unter Kostengesichtspunkten der wirtschaftlichere Weg, kann es auch für Teile des digitalen Bestands einer Einrichtung sinnvoll sein, diese durch den Service eines solchen Anbieters archivieren zu lassen. Eine andere Möglichkeit bietet sich in dem beschriebenen Hardware-Hosting bzw. Storage-Betrieb durch einen ausgewiesenen Dienstleister.

Quellen und Literatur

Borghoff, Uwe M. [u.a.] (Hrsg.): Long-Term Preservation of Digital Documents : Principles and Practices. Heidelberg [u.a.] : Springer, 2003

Kopal (2007): kopal: Ein Service für die Langzeitarchivierung digitaler Informationen. Projekt kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen), 2007 (s. http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf)

Lutterbeck, Bernd / Bärwolff, Matthias / Gehring, Robert A. (Hrsg.): Open Source Jahrbuch 2007 : Zwischen freier Software und Gesellschaftsmodell. Berlin : Lehmanns Media, 2007 (s. http://www.opensourcejahrbuch.de/download/jb2007)

6 Organisation

Christian Keitel

Einführung

Die Organisation der digitalen Langzeitarchivierung kann als die arbeitsteilige Bewältigung dieser Aufgabe verstanden werden. Die bisherigen Erfahrungen zeigen, dass die bei der digitalen Langzeitarchivierung anstehenden (Teil-)Aufgaben sehr unterschiedlich organisiert und voneinander abgegrenzt werden können. Zumeist erfolgt eine Aufgabenteilung zwischen einzelnen, mit der Langzeitarchivierung befassten Institutionen. Modelle zur Arbeitsteilung innerhalb einer Institution (interne Organisationsmodelle) werden kaum veröffentlicht, zumal die Aktivitäten derzeit oft nur einen zeitlich befristeten Projektstatus besitzen. Eine allgemeine und generalisierende Darstellung der Organisation ist daher nur auf einem sehr abstrakten Niveau möglich. Einen solchen Rahmen bietet das Funktionsmodell (*functional modell*) von OAIS. Anschließend an diese aufgabenorientierte Beschreibung werden weitere Faktoren genannt, die bei der Organisation der digitalen Langzeitarchivierung zu berücksichtigen sind. In einem dritten Schritt werden anhand konkreter Beispiele mögliche Umsetzungen skizziert.

1. Aufgaben nach dem OAIS-Funktionsmodell

Produktion: Die Entstehung der Daten ist nach OAIS nicht Bestandteil eines digitalen Langzeitarchivs. Aus diesem Grund müssen im Bereich *Ingest* Schnittstellen und Übergabe- bzw. Aufnahmeprozeduren detailliert beschrieben werden. In den meisten Fällen ist das digitale Archiv organisatorisch vom Produzenten getrennt. Manchmal wird diese Trennung aber auch relativiert oder aufgehoben:

Archivierung durch die Produzenten (1): 1996 wurde den australischen Behörden nach der Theorie des *records continuum* auferlegt, alle alten, im Dienst nicht mehr benötigten, Dokumente dauerhaft selbst zu verwahren. Den Archiven kam dabei die Rolle zu, das Funktionieren des Konzepts sicherzustellen, also eine Art „Archivierungspolizei" zu spielen. Bereits 2000 kehrte das Australische Nationalarchiv wieder zu seiner traditionellen Politik zurück, d.h. zur Übernahme dieser Dokumente. Nur Archive und Bibliotheken haben ein genuines Interesse an der Erhaltung von Informationen, die in den Augen ihrer Ersteller „veraltet" sind. Erst dieses Interesse gewährleistet, dass vermeintlich uninteressante Daten weiterhin gepflegt werden.

Archivierung durch die Produzenten (2): Die Systeme der Umweltbeobachtung verwahren aktuell produzierte Daten zusammen mit den Daten vergangener Jahrzehnte. Die einzelnen Informationen sollen dauerhaft im selben System und unter denselben Namen aufgefunden und angesprochen werden, die systemische Einheit dieser Daten ist über einen langen Zeitraum hinweg erwünscht. Die Information veraltet also im Gegensatz zum beschriebenen australischen Beispiel theoretisch nie. Vergleichbare Systeme werden derzeit in vielen Naturwissenschaften aufgebaut. Ist es nicht ganz allgemein sinnvoll, bei der Entstehung der Daten dieselben Erhaltungsregeln anzuwenden wie später im Archiv? Analog hierzu sieht sich das 2003 gegründete britische *Digital Curation Centre* auch für den ganzen Lifecycle eines Dokuments zuständig: "The term 'digital curation' is increasingly being used for the actions needed to maintain and utilise digital data and research results over their entire life-cycle for current and future generations of users."[1]

Archivisches Engagement bei den Produzenten: Seit über 15 Jahren engagieren sich die klassischen Archive in den Behörden bei der Einführung elektronischer Akten und anderer digitaler Systeme. Ihr Motiv: Bei der Einführung eines Systems werden die Grundlagen dessen gelegt, was

1 JISC Circular 6/03 (Revised), in: http://www.dcc.ac.uk/docs/6-03Circular.pdf.

dann später im Archiv ankommt. Danach ist es weniger aufwändig, in der Behörde Dinge grundsätzlich zu regeln, als später jedes Objekt einzeln nachbearbeiten zu müssen. Im DOMEA-Konzept (**Do**kumenten**m**anagement und **E**lektronische **A**rchivierung) werden die beiden Bereiche auch begrifflich zusammengezogen.

Archive werden zu Produzenten: Durch die Digitalisierungsstrategien der Archive und Bibliotheken mutieren diese klassischen Gedächtnisinstitutionen auf einmal selbst zu Datenproduzenten. Es bedarf zwar zusätzlicher Qualitätssicherungsmaßnahmen für die Digitalisate, eine Ingest-Schnittstelle oder die Umwandlung von SIPs in AIPs sind jedoch nicht mehr notwendig.

Ingest: Setzt man mit OAIS eine Trennung zwischen Produktions- und Archivzuständigkeit, dann müssen im Ingest die Übernahmepakete (SIPs) entgegen genommen, überprüft, und in Archivierungspakete (AIPs) umgewandelt werden. Beschreibende Metadaten werden extrahiert und an das Data Management weitergegeben. Der ebenfalls von den OAIS-Autoren verfasste PAIMAS-Standard gliedert diesen Bereich in insgesamt vier Phasen: Nach einer Vorbereitungsphase werden in einer Definitionsphase alle wesentlichen Rahmenbedingungen vereinbart und erprobt. Hierzu gehört insbesondere die Auswahl der zu übernehmenden Objekte und die Abklärung sämtlicher rechtlichen Aspekte. Während der Transferphase werden diese übernommen und schließlich in der Validierungsphase auf ihre angenommenen Eigenschaften hin überprüft.
Auch bei einer festen Trennung zwischen Produzenten und Archiv können die einzelnen Aufgaben sehr unterschiedlich aufgeteilt werden. Hierzu gehören die Auswahl der Objekte, ihre Ausstattung mit Metadaten und die ggf. erforderliche Migration der Objekte in ein archivierungsfähiges Format. Entsprechend kann sich die dem Archiv verbleibende Ingest-Aufgabe v.a. administrativ gestalten (d.h. es gibt dem Produzenten die entsprechenden Vorgaben) oder zunehmend auch technische Komponenten enthalten (d.h. es setzt diese Punkte selbst um). Die Entscheidung für eine der beiden Optionen ist wesentlich von der Gleichartigkeit der Objekte abhängig: Erst wenn sich die Objekte sehr stark gleichen, kann die Zahl der Vorgaben so weit reduziert werden, dass eine entsprechende Automatisierung auch erfolgreich umgesetzt werden kann. Bei stark differierenden Objekten lassen sich diese Regeln nicht in einer vergleichbar umfassenden Weise aufstellen, weshalb die Aufgaben vom Archiv selbst übernommen werden müssen, was dessen Aufwand entsprechend erhöht.
Im letztgenannten Fall können dann weitere Teilaufgaben gebildet werden. Bei-

spielsweise kann die Metadatenerfassung in zwei aufeinanderfolgende Schritte aufgespalten werden: a) Anlegen erster identifizierender Metadaten. b) Nähere Beschreibung im Zuge der weiteren Bearbeitung.

Archival Storage: In diesem Bereich werden die AIPs über einen langen Zeitraum gespeichert. Der Zustand der Speichermedien wird kontinuierlich überwacht, im Bedarfsfall werden einzelne Medien ersetzt, regelmäßig werden auch ganze Medien-Generationen in neuere Speichertechnologien migriert. Neben Hardware und Software sind hier also v.a. IT-Kenntnisse erforderlich. Es ist daher auch der Bereich, der am ehesten von den klassischen Gedächtnisinstitutionen an externe Rechenzentren ausgelagert wird. Andererseits unterscheiden sich die Anforderungen der digitalen Langzeitarchivierung z.T. erheblich von denen, die gewöhnlich an Rechenzentren gestellt werden. Die National Archives and Records Administration (NARA) der Vereinigten Staaten hat daher Anfang der 1990er Jahre den Bereich wieder ins Haus geholt.

Data Management: In diesem Bereich werden die identifizierenden, beschreibenden und administrativen Metadaten gepflegt. Er ist daher für die klassischen Gedächtnisinstitutionen nicht neu. Sofern nicht ein eigenes Recherchesystem für die digitalen Objekte aufgebaut wird, liegt es nahe, dass dieser Bereich von den Organisationseinheiten übernommen wird, die bereits für die Beschreibung der analogen Objekte zuständig sind.

Preservation Planning: Digitale Langzeitarchivierung erfordert eine kontinuierliche aktive Begleitung der archivierten Objekte. Zentral ist die Terminierung und Koordination der einzelnen Erhaltungsprozesse. Schnittstellen bestehen zu den Bereichen Ingest, Archival Storage und Data Management.

Access: Diese Einheit ermöglicht die Benutzung der digitalen Objekte. Sie ermöglicht die Recherche in den beschreibenden Metadaten und liefert die Benutzungspakete aus (DIPs). Manche Archive überlassen diese Aufgabe aber auch ihren Benutzern, d.h. ausgegeben werden die nicht weiter veränderten AIPs.

Administration: Der Bereich klärt das Zusammenspiel der einzelnen Organisationseinheiten. Er handelt grundsätzliche Vereinbarungen mit den Produzenten aus, definiert die Rahmenbedingungen für eine Benutzung, überwacht das Archivsystem, entwickelt Standards und Policies und berichtet regelmäßig dem außerhalb des OAIS angesiedelten Management. Er ist somit kaum technisch geprägt.

2. Weitere Faktoren

Die Organisation der digitalen Langzeitarchivierung ist außer von den Aufgaben und den zu archivierenden Objekten auch von weiteren Faktoren abhängig. Genannt werden können die Größe der Einrichtung, ihre sonstigen Aufgaben und die Qualifikation ihres Personals. Sehr große Archive können zu jeder Einheit des OAIS-Funktionsmodells mindestens eine administrative Einheit bilden. Zusätzlich kann noch ein Forschungsbereich ausgegliedert werden. Kleinere Archive sind dagegen gezwungen, mit weniger administrativen Einheiten auszukommen. Bei klassischen Gedächtniseinrichtungen stellt sich die Frage, welche Aufgaben unabhängig vom Medientyp bearbeitet werden können. In zahlreichen Bereichen sind zudem sowohl die Kenntnisse traditionell ausgebildeter Archivare oder Bibliothekare als auch ausgeprägte IT-Kenntnisse erforderlich. Die Organisation ist daher auch von dem bereits bestehenden Personalbestand der Einrichtung und der Möglichkeit einer Neueinstellung abhängig.

3. Beispiele/Umsetzung in die Praxis

3.1. Centre national d'études spatiales (CNES)

Die französische Raumfahrtagentur CNES archiviert fast ausschließlich digitale Daten. Es wurden drei administrative Einheiten gebildet: a) Ingest, b) Archival Storage und c) Data Management und Access. Im Ingest arbeiten Archivare und Computerspezialisten zusammen. Der Archivar definiert die zu übernehmenden Objekte, überprüft sie auf ihre Vollständigkeit und strukturiert sie. Der Computerspezialist definiert Daten und Metadaten, nimmt die physische Übernahme und die Validierung vor und entwickelt entsprechende Tools. Das neue Berufsbild des *Digital Data Manager* kann auf beiden Gebieten des Ingest tätig werden. Beim Archival Storage werden ausschließlich Computerspezialisten eingesetzt. Seit 1994 wird dieser Bereich vom STAF (*Service de Transfert et d'Archivage de Fichiers*) ausgeführt. Die OAIS-Bereiche Data Management und Access werden beim CNES zusammengezogen. Im Vordergrund stehen Datenbank-, Retrieval- und Internettechnologien, daneben werden vertiefte Kenntnisse über das Archiv benötigt. Das Funktionieren des Archivs wird durch eine Koordinationsstelle, bewusst klein konzipierte Überlappungsbereiche und die weitgehende Unabhängigkeit der einzelnen Einheiten gewährleistet.

3.2 The National Archives (UK)

Die National Archives haben mehrere objektspezifische Ansätze zur digitalen Archivierung entwickelt, die zusätzlich von zentralen Systemen (z.B. die Formatdatenbank *PRONOM*) unterstützt werden. Seit 2001 ist zudem für die Erhaltung von *born digital material* nicht mehr das *Records Management Department* sondern das neu eingerichtete *Digital Preservation Department* zuständig. Für strukturierte Daten wurde 1997 eine Kooperationsvereinbarung mit dem Rechenzentrum der Londoner Universität (University of London Computer Centre, UCLL) geschlossen, in deren Folge das *National Digital Archive of Datasets* (NDAD) 1998 in Betrieb genommen werden konnte. Die National Archives sind für die Auswahl der Daten und die Definition der Service-Levels zuständig, NDAD für alle weiteren Aufgaben (explizit unterschieden werden Ingest, Preservation und Access). Im NDAD arbeiten zwölf Personen in vier Disziplinen: Die *Project Archivists* treffen zentrale Entscheidungen über die Organisation des Archivs, Katalogisierung und Indexierung und leiten die Computer-Spezialisten an. Die *Archive Assistants* sind für die Benutzerbetreuung zuständig. Sie unterstützen die Project Archivists z.B. durch Einscannen der Papierdokumentation. Die *Data Specialists* sind für die technische Umsetzung der getroffenen Entscheidungen zuständig. Der *Systems Support Staff* stellt schließlich das Funktionieren von Hard- und Software sicher. Für die Archivierung elektronischer Records (Akten) wurde in den National Archives Mitte der 1990er Jahre das EROS-Projekt aufgesetzt, das nun im Seamless-Flow-Programm fortgesetzt wird. Erste Ergebnisse sind ab Ende 2007 zu erwarten. Gleichzeitig werden im 2003 in den National Archives gegründeten *Digital Archive* bereits Records übernommen und Erfahrungen aufgebaut. Für die Archivierung von Internetseiten haben sich die National Archives 2003 mit der British Library, den Nationalbibliotheken von Wales und Schottland, JISC und dem Wellcome Trust zum *UK Web Archiving Consortium* zusammengeschlossen, um eine gemeinsame Infrastruktur zur Web-Archivierung aufzubauen[2].

3.3 Deutsche Nationalbibliothek (DNB) und Staats- und Universitätsbibliothek Göttingen (SUB)

Die Deutsche Nationalbibliothek und die Staats- und Universitätsbibliothek Göttingen haben ihre Lösung zur Archivierung digitaler Objekte im Projekt

2 Auf einer vergleichbaren Kooperation basiert das BOA-Projekt. Die beiden Landesbibliotheken und das Landesarchiv von Baden-Württemberg sind zuständig für die Auswahl und den Ingest der Webseiten und Einzel-Dokumente, während das Bibliotheksservicezentrum Baden-Württemberg das Speichersystem und die Infrastruktur zur Verfügung stellt.

KOPAL gemeinsam mit der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und der IBM Deutschland entwickelt. Die Partner gehen von einem arbeitsteiligen Vorgehen aus: Die Übernahme und Aufbereitung der AIPs liegt in den Händen der beteiligten Bibliotheken und erfolgt durch eine OpenSource-Software. Die fertigen AIPs werden dann per Fernzugriff zentral im Rechenzentrum der GWDG gespeichert. Dabei kommt das durch die IBM entwickelte DIAS-System zu Einsatz. Die Benutzung erfolgt dann wiederum durch Fernzugriff bei den beiden Bibliotheken.

Literatur

- Reference Model for an Open Archival Information System (OAIS), Blue Book 2002, http://www.ccsds.org/publications/archive/650x0b1.pdf
- Producer-Archive Interface Methodology Abstract Standard, Blue Book 2004, http://public.ccsds.org/publications/archive/651x0b1.pdf (PAIMAS)
- DOMEA-Konzept: Das Organisationskonzept, die Erweiterungsmodule und weitere Informationen finden sich auf den Seiten der KBSt: http://www.kbst.bund.de.
- Adrian Brown, Developping practical approaches to active preservation, in: Proceedings of the 2nd International Conference on DIgital Curation, Glasgow 2006
- Adrian Brown, Archiving Websites. A Practical Guide for Information Management Professionals, London 2006
- Claude Huc, An organisational model for digital archive centres, http://www.erpanet.org/events/2004/amsterdam/presentations/erpaTraining-Amsterdam_Huc.pdf; auch als ...ppt und ...m3u.
- Richard Jones, Theo Andrew, John MacColl, The Institutional Repository, Oxford 2006
- Patricia Sleeman, It's Public Knowledge: The National Digital Archive of Datasets, in: Archivaria 58 (2004), S. 173 - 200

7 Das Referenzmodell OAIS - Open Archival Information System

Nils Brübach

[überarbeitete Fassung eines Vortrags, gehalten auf der 6. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen" am 5./6. März 2002 in Dresden]

Bearbeiter: Manuela Queitsch, Hans Liegmann

Das als ISO 14721 verabschiedete Referenzmodell „Open Archival Information System – OAIS" beschreibt ein digitales Langzeitarchiv als eine Organisation, in dem Menschen und Systeme mit der Aufgabenstellung zusammenwirken, digitale Informationen dauerhaft über einen langen Zeitraum zu erhalten und einer definierten Nutzerschaft verfügbar zu machen.
Im folgenden Beitrag werden vier Ziele verfolgt: Erstens sollen die Entwicklung des OAIS, sein Konzept und sein Ansatz skizziert werden. Zweitens werden die wesentlichen Kernkomponenten des OAIS, nämlich die in ihm vorgesehenen Informationsobjekte bzw. Informationspakete und das ihnen zu Grunde liegende Datenmodell analysiert und vorgestellt, um drittens das Funktionsmodell des OAIS zu erläutern. Es ist ein besonderes Kennzeichen des OAIS, das bereits bei seiner Entwicklung nicht nur ein auf theoretischen Überlegungen fußendes Modell entwickelt wurde, sondern das die Frage nach der Anwend-

barkeit und deren Prüfung vorab an konkreten Anwendungsfällen mit in die Konzeption und Entwicklung einbezogen wurden. Deswegen wird im vierten Abschnitt kurz auf einige bereits existierende Anwendungsbeispiele des OAIS eingegangen: OAIS ist kein am „grünen Tisch" auf Basis rein theoretischer Überlegungen entwickelter Ansatz, sondern für die Praxis entwickelt worden.

1. Die Entwicklung des OAIS und sein Ansatz

Das Open Archival Information System hat seine Wurzeln im Gebiet der Raumfahrt. Diese Tatsache ist nur auf den ersten Blick wirklich überraschend, wird aber verständlich, wenn man sich vor Augen führt, dass in diesem Bereich seit den sechziger Jahren elektronische Daten in großen Mengen angefallen sind demzufolge die das klassische öffentliche Archivwesen jetzt beschäftigenden Fragen schon weit eher auftreten mussten. Federführend für die Entwicklung des OAIS, die seit dem Jahre 1997 betrieben wurde, war das „Consultative Committee for Space Data Systems", eine Arbeitsgemeinschaft verschiedener Luft- und Raumfahrtorganisationen wie etwa der NASA oder der ESA oder der Deutschen Gesellschaft für Luft- und Raumfahrt unter Federführung der NASA. Beteiligt waren von archivischer Seite seit 1997 die amerikanische nationale Archivverwaltung (NARA) und die Research Libraries Group (RLG). Das OAIS wurde im Jahre 1999 erstmals als vollständige Textfassung in Form eines so genannten „Red Book" vorgelegt[1]. Lou Reich und Don Sawyer von der CCSDS bzw. der NASA sind die Autoren der unterschiedlichen Textfassungen und hatten auch die Koordination der Arbeitsgruppe zur Erstellung des Textes inne. Im gleichen Jahr 1999, in dem das Red Book veröffentlicht und der internationalen Fachgemeinschaft der Archivarinnen und Archivare zur Diskussion gestellt wurde, wurde die Vorlage auch bei der ISO als internationaler Standard eingereicht. Er durchlief dort die üblichen Prüfungsverfahren. Der Text des Red Book wurde nach Ergänzung und Überarbeitung im Juni 2001 als ISO/DIS 14721 angenommen und zum 1. Januar 2002 in das Normenwerk der Internationalen Standardorganisation integriert; die Übernahme in das deutsche Normenwerk steht allerdings noch aus. Wir haben es also für diesen Bereich, ähnlich wie bei der ISO/DIN 15489 „Schriftgutverwaltung", erneut mit einem Standard zu tun und nicht nur mit einem Arbeitsdokument unter vielen. Allein schon das Abstimmungsverfahren und die nur wenigen vorgenommenen Än-

1 http://public.ccsds.org/publications/archive/650x0b1.pdf. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2003.

derungen am ursprünglichen Text des Red Book zeigen, wie ausgefeilt und wie weit entwickelt das Projekt bereits gediehen war, als es bei der ISO als Standard vorgelegt wurde. Dieses Arbeitsverfahren - mit Hilfe von Standards gesicherte Arbeitsergebnisse zu einer Art von „anwendungsbezogenem Allgemeingut" werden zu lassen - scheint sich im Bereich der Archivierung elektronischer Unterlagen immer stärker durchzusetzen: So wurde vom ISO TC 46 und TC 171 eine Untermenge des PDF-Formats (PDF/A = PDF/Archive) ein Standardisierungsprozess (ISO 19005-1. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF (PDF/A)) eingeleitet, der zur größeren Akzeptanz des Formats für die Langzeitarchivierung digitaler Dokumente führen soll.[2]

Das OAIS-Konzept ist ein Standard in Form eines Referenzmodells für ein dynamisches, erweiterungsfähiges Archivinformationssystem. Ganz bewusst versteht sich OAIS als ein offener Standard, aber auch als ein Modell, das den Anspruch der Allgemeingültigkeit verfolgt. Das hat zwei Konsequenzen:

- erstens verzichtet OAIS auf eine Beschränkung auf bestimmte Datentypen, Datenformate oder Systemarchitekturen (im technischen Sinne) und
- zweitens will OAIS anwendungsfähig und skalierbar sein für eine Vielzahl bestimmter Institutionen und ihre jeweiligen Bedürfnisse. Der Text des OAIS hat insgesamt sieben Abschnitte.

Kapitel 1 „Einführung" beschreibt die Zielsetzung, den Anwendungsrahmen, bestimmte Anwendungsregeln und stellt üblicherweise die notwendigen Begriffsdefinitionen voran.

In Kapitel 2 wird das Konzept des OAIS, d.h. die unterschiedlichen Typen von Informationen, die modellierbaren standardisierten Operationen und auch die Systemumgebung (im funktionalen Sinne) beschrieben.

Kapitel 3, eines der Kernkapitel, beschreibt die Tätigkeitsfelder eines OAIS-Betreibers.

Kapitel 4 ist den Datenmodellen gewidmet, die dem OAIS zugrunde liegen. Hier wird einerseits das Funktionsmodell beschrieben und andererseits die unterschiedlichen Informationspakete, ihre Verwendung und ihre Verknüpfung zu einem Informationsmodell.

Kapitel 5 ist der zweite Kernbereich, denn hier wird beschrieben, welche Operationen für eine dauerhafte Aufbewahrung digitaler Aufzeichnungen und für

2 Der Begriff „Langzeitarchivierung" wird als Äquivalent zum englischen Terminus „long-term preservation" verwendet. Er ist als technischer Begriff zu sehen, der darauf hin deuten solle, dass anders als bei der Archivierung im analogen Umfeld, die dauerhafte Aufbewahrung von digitalen Objekten eben nicht auch die dauerhafte Zugänglichkeit automatisch nach sich zieht.

die Gewährleistung des Zugangs zu ihnen unverzichtbar sind. Die heutige Archivlandschaft ist eine offene Archivlandschaft.
Demzufolge widmet sich Kapitel 6 dem Betrieb eines Archivs nach OAIS-Standard in Kooperation mit anderen Archiven. So entscheidende Fragen wie die der technischen Kooperation, die Frage nach Funktion und Aufbau von Schnittstellen und eines gemeinsamen übergreifenden Managements verschiedener digitaler Archive werden hier angesprochen.
Der 7. Teil des Standards enthält die Anhänge, in denen Anwendungsprobeläufe beschrieben werden, auf andere Standards verwiesen wird, Modelle für Kooperationen skizziert und Entwicklungsmodelle für bestimmte Software-Lösungen zumindest angedeutet werden.[3] Auf diesen letzen Aspekt der „Interoperabilität" sei an dieser Stelle besonders hingewiesen. OAIS versteht sich nämlich nicht als eine geschlossene Lösung, sondern als ein offenes Informationssystem, das in jedem Fall und in jedem Stadium mit anderen Parallelsystemen vernetzbar sein soll. Dadurch, dass OAIS sich selbst als Referenzmodell definiert, ist es auch offen für verschiedene technische Lösungsmöglichkeiten, die aber über den zentralen Punkt der funktionalen Interoperabilität aufeinander abgestimmt und miteinander verknüpfbar sein müssen.
Das Open Archival Information System beschreibt ein Informationsnetzwerk, das den Archivar und den Nutzer als Hauptkomponenten des digitalen Archivs versteht. Archivierung ist nicht an Maschinen delegierbar: Der Mensch hat im Sinne des OAIS die Verantwortung für die Sicherung von Informationen und deren Bereitstellung für eine bestimmte Nutzergruppe. Die Unterscheidung verschiedener Nutzergruppen (Designated Communities) ist eine Besonderheit des OAIS. Die Interoperabilität liegt nämlich nicht nur in technischer und in funktioneller Ebene, sondern eben auch darin, dass unterschiedliche Benutzergruppen unterschiedliche Anforderungen an elektronische Archive in der Gegenwart haben und in der Zukunft haben werden: Anforderungen, die heutige Entwicklergenerationen technischer Lösungen überhaupt nicht voraussehen können und bei denen das, was Archivierung eigentlich ausmacht - Sicherung von Authentizität und Integrität durch dauerhafte Stabilisierung und Zugänglichmachung von authentischen unikalen Kontexten - auch im digitalen Umfeld gewährleistet ist. Die Offenheit des OAIS ist also auf Zukunftsfähigkeit und auf Nachhaltigkeit ausgerichtet. Die heute im Rahmen des OAIS realisierten Lösungen sollen auch in der Zukunft verwendbar und in neue technische Realisierungen übertragbar sein. Das OAIS wird damit auch offen für neue Anforderungen an die Nutzung.

3 Gail M. Hogde: Best Practices for Digital Archiving. In D-LIB-Magazine, Vol.6 No.1, January 2000, S.8. [http://www.dlib.org/dlib/january00/01hodge.html]

Das OAIS konzentriert sich auf die Langzeitaufbewahrung und Langzeitnutzbarhaltung hauptsächlich digitaler Aufzeichnungen und dies unter Berücksichtigung der sich verändernden Technologien. Wenn die Autoren des OAIS sich hauptsächlich auf digitale Aufzeichnungen konzentrieren, so verweisen sie doch darauf, dass in einem weiteren Sinne jedes digitale Archiv, das dem OAIS-Standard folgt, immer auch mit schon bestehenden, sich auf analoge Unterlagen konzentrierenden Archivlösungen verknüpfbar sein und dass diese Verknüpfung auch in der Zukunft erhalten bleiben muss. Das OAIS zeigt also Wege auf zur dauerhaften Sicherung digitaler Unterlagen in ihrem Kontext und den wechselseitigen Beziehungen zu analogem Schriftgut, die sich wandeln können: Die Gedächtnisorganisationen werden in Zukunft eben auch Papier enthalten müssen, es treten neue Aufzeichnungsformen hinzu, die die alten keineswegs vollständig verdrängen werden. Ebenso wie sich das noch vor wenigen Jahren propagierte „papierlose Büro“ als Hirngespinst erwiesen hat und, viel bescheidener, heute nur noch vom „papierarmen Büro“ gesprochen wird, sind Überlegungen zu einem vollständigen Medienbruch bei der Archivierung realitätsfremd. Das OAIS berücksichtigt Bestehendes: Es ist gerade deshalb ein Modellansatz und ein Standard, der damit auch Einfluss auf zukünftige Arbeitsmethoden im Archiv nehmen wird. Es geht nämlich von den klassischen archivischen Arbeitsfeldern, Erfassen, Aussondern, Bewerten, Übernehmen, Erschließen, Erhalten und Zugänglichmachen aus, aber definiert sie in ihren Teilaufgaben und Arbeitsabläufen unter dem Blickwinkel der Bedürfnisse digitaler Archivierung neu. Im gewissen Sinne beantwortet der Text des OAIS die schon so häufig gestellte, aber bisher bestenfalls unbefriedigend beantwortete Frage nach dem zukünftigen Aufgabenspektrum von Gedächtnisorganisationen im digitalen Zeitalter. Auch die Frage danach, welche Funktionen automatisierbar sind, wird thematisiert. Hier liegt nicht zuletzt auch ein für Fragen der Aus- und Fortbildung interessanter Aspekt.
Das OAIS erhebt den Anspruch, auf jedes Archiv anwendbar zu sein, Archiv vom Begriff her bezieht sich hier ausdrücklich auf den Bereich der dauerhaften Aufbewahrung und langfristigen Zugangssicherung. Dabei wird auch kein Unterschied gemacht, ob die Archivierung organisationsintern bei den produzierenden Stellen selbst erfolgt, oder bei Organisationen, die digitale Objekte zur Archivierung übernehmen.

2. Die Kernkomponenten: Informationsobjekte und Datenmodell

Das OAIS unterscheidet zwischen drei so genannten Informationsobjekten die miteinander in Verbindung stehen und sich aufeinander beziehen, aber entwi-

ckelt worden sind, um den unterschiedlichen Umgang und die unterschiedlichen Tätigkeiten bei der digitalen Archivierung besser beschreiben zu können. Das was Archive an digitalen Unterlagen übernehmen, heißt in der Terminologie des OAIS Submission Information Packages (SIP). Im Archiv selbst werden diese SIP vom Archiv durch Metainformationen ergänzt und umgeformt zu Archival Information Packages (AIP), die weiter verarbeitet werden und die im Kern die Form darstellen, in der die digitalen Informationen tatsächlich langfristig aufbewahrt werden. Zugänglich gemacht werden die AIPs über die so genannten Dissemination Information Packages (DIP), die für bestimmte Nutzergruppe je nach Vorliegen bestimmter rechtlicher Bedürfnisse generiert und zielgruppenorientiert zur Verfügung gestellt werden können. Dieser Ansatz ist im Vergleich zum klassischen Bestandserhaltung durchaus ungewöhnlich. Im Sinne des OAIS wird nämlich nicht ohne Veränderung das einfach aufbewahrt, was man übernimmt, sondern es wird zukünftig die Aufgabe der Verantwortlichen sein, sehr viel mehr noch als im Bereich der Archivierung von analogen Unterlagen dafür zu sorgen, dass die Unterlagen überhaupt archivfähig sind. Die Umformung der SIPs zu Archival Information Packages kann z.B. darin bestehen, dass aus den mit übernommenen Objekten und den mitgelieferten Metadaten die zur Langzeiterhaltung notwendigen Metadaten generiert werden. Darüber hinaus sind die Formate, in denen ein SIP dem Archiv angeboten und von ihm übernommen wird, keinesfalls unbedingt identisch mit den tatsächlichen Aufbewahrungsformaten, in denen die Archival Information Packages dann tatsächlich vorliegen. Sichergestellt sein muss die Bewahrung von Authentizität und Integrität auch mit Blick auf die rechtswahrende und rechtssichernde Funktion digitaler Archive. Ein AIP aus dem Jahre 2003 wird naturgemäß in einem ganz anderen Format und in einer ganz anderen Datenstruktur vorliegen, als das gleiche AIP etwa im Jahre 2010. Grundgedanke dieser Arbeit mit Informationspaketen ist es, dass Inhalte, Metadaten und - wo unverzichtbar - die entsprechenden Strukturen der digitalen Aufzeichnungen nachvollziehbar bzw. rekonstruierbar gehalten werden, unabhängig von den sich wandelnden technischen Gegebenheiten. Dies ist ein Aspekt, der eben auch auf die Benutzung der Unterlagen zielt. Die Dissemination Information Packages dienen der Nutzung und dem Zugang je nach den Bedürfnissen der jeweiligen Benutzergruppen und sind ganz gezielt für unterschiedliche Benutzer anzupassen und auch anpassbar zu erhalten. Gerade das ist für die klassische dauerhafte Bestandserhaltung in Archiven eine ungewöhnliche Vorstellung: dem Benutzer wird nicht mehr das vorgelegt, was im Magazin verwahrt wird, sondern aus dem was verwahrt wird werden Informationspakete generiert, die auf die Bedürfnisse der Kunden natürlich auch in Abhängigkeit von die Nutzung einschrän-

kenden Rechten Betroffener oder Dritter zugeschnitten werden. Diese Umformung der AIPs in DIPs bezieht sich dabei keinesfalls ausschließlich auf die Veränderung der Datenformate, sondern eben auch auf die Bereitstellung von digitalen Informationen in Verbindung mit einer für den Benutzer besonders komfortablen Funktionalität. Hier wird im OAIS ein Ansatz aufgegriffen, der im Bereich der archivischen online-Findmittel verwendet wird. Die einzelnen Informationspakete werden im Rahmen des OAIS als digitale Objekte verstanden. Sie bestehen immer aus Daten und beschreibenden und ggf. ergänzenden, repräsentativen Zusatzinformationen.
Jedes Informationspaket enthält erstens inhaltliche Informationen (Content Information), die aus den übernommenen, ggf. aufbereiteten Ursprungsdaten und der beschreibenden Repräsentationsinformation bestehen, und zweitens so genannte „Informationen zur Beschreibung der Aufbewahrungsform" (Preservation Description Information (PDI)), die erklären, was an Technik und welche Verfahren auf die Inhaltsinformation angewandt wurden, also wie sie verändert wurden und welche Technik und welche Verfahren benötigt werden, um sie zu sichern, sie eindeutig zu identifizieren, sie in ihren Kontext einzuordnen und für die Zukunft nutzbar zu machen. Die Preservation Description enthält Informationen, die die dauerhafte Aufbewahrung beschreibt, sie besteht wiederum aus vier Elementen.
Erstes Element ist die Provenienz, hier werden also die Quelle der Inhaltsinformation seit deren Ursprung und ihre weitere Entwicklung, also ihr Entstehungs- und Entwicklungsprozess, beschrieben.
Zweites Element ist der Kontext, wo die Verbindung einer konkreten Inhaltsinformation mit anderen Informationen außerhalb des jeweiligen Informationspakets nachvollziehbar gehalten wird.
Drittes Element sind Beziehungen (References), wo über ein System von eindeutigen Bezeichnern (unique identifiers) die Inhaltsinformationen mit den auf sie bezogenen Metadaten und anderen Inhaltsinformationen eindeutig identifizierbar und eindeutig unterscheidbar gemacht werden.
Viertes Element sind Informationen zur Stabilisierung (fixity), damit die Inhaltsinformationen vor nicht erfasster Veränderung bewahrt werden können.

3. Das Funktionsmodell des OAIS

Es sind sechs Aufgabenbereiche[4], die im Rahmen des skizzierten Standards beschrieben werden:

4 Vgl. Grafik 7.1

1. Datenübernahme (Ingest)
2. Datenaufbewahrung (Archival Storage)
3. Datenmanagement
4. Systemverwaltung
5. Planung der Langzeitarchivierung (Preservation Planning)
6. Zugriff (Access)

Im Bereich Ingest geht es um die Übernahme des digitalen Archivguts. Zunächst wird die Vorbereitung der Einlagerung im Archiv vorzunehmen sein, dazu gehört etwa auch die Bereitstellung der notwendigen technischen Kapazitäten und die Kontaktaufnahme mit dem Produzenten. Ein weiterer Aspekt, der ganz entscheidend ist, ist die Qualitätssicherung der Submission Information Packages, d.h. ihre Prüfung auf Lesbarkeit, Verständlichkeit und korrekten Kontext und dann die Herstellung der archivischen Informationspakete (AIP), die mit den Formaten und Standards des jeweils aufbewahrenden Archivs übereinstimmen. Der Analyse, Sicherung und ggf. Verbesserung der Datenqualität kommt im digitalen archivischen Vorfeld eine Schlüsselrolle zu, hier wird aber auch erstmalig verändernd eingegriffen. Das OAIS geht davon aus, dass digitale Archive aus ganz unterschiedlichen Systemumgebungen SIPs in einer Vielzahl von unterschiedlichen Formaten einfach übernehmen müssen und diese erst bei der digitalen Archivierung, also bei der Einlagerung ins digitale Magazin, zu nach einheitlichen Standards aufgebauten und zu generierenden AIPs umformen. Zum Bereich Übernahme gehört auch die Erstellung der notwendigen Erschließungsinformationen für die Erschließungsdatenbank des digitalen Archivs und erste planende Maßnahmen, die das regelmäßige Update des Datenspeichers und das dazu notwendige Datenmanagement organisieren.
Der zweite Teil „Achival Storage“ umfasst den digitalen Speicher, seine Organisation und seinen Aufbau im engeren Sinne. Hier werden die AIPs vom Übernahmebereich in Empfang genommen und eingelagert und es wird dafür gesorgt, dass regelmäßig gewartet und die Wiederauffindbarkeit der archivischen Informationspakete überprüft wird. Dazu gehört der Aufbau einer technischen Lagerungshierarchie und die regelmäßige systematische Erneuerung der im jeweiligen Archiv standardisiert verwendeten Datenträger, sowie das so genannte Refreshing, d.h. die Überprüfung der verwendeten Datenträger auf ihre Lesbarkeit und die Verständlichkeit der gespeicherten AIP. In diesem Zusammenhang ist darauf zu verweisen, das OAIS ausdrücklich die Vorteile einer redundanten Archivierung auf zwei verschiedenen Informationsträgern hervorhebt.
Im Bereich Datenmanagement geht es um die Wartung und das Zugänglichhal-

ten der Verzeichnungsinformationen und ihre kontinuierliche Ergänzung und Aufbereitung, dann aber auch das Verwalten verschiedener Archivdatenbanken und auch in diesem Bereich die Ausführung von verschiedenen Datenbank-Updates zur Sicherung von Lesbarkeit, Verständlichkeit und Nutzbarkeit.
Punkt vier umfasst das Management des OAIS. Management bezieht sich auf die Beziehungen zwischen Archivaren und Nutzern auf der einen Seite und dem Software/Hardware-System auf der anderen. Beschrieben werden alle Regelungen zur Zuständigkeit für die Arbeitsvorgänge im Archivssystem, wozu auch gehört, dass das, was automatisierbar ist, von den Vorgängen getrennt wird, die von Menschen erledigt werden müssen. Ebenso der Bereich Qualitätssicherung ist hier eingeordnet. Auch das Aushandeln von Verträgen zur Übergabe und zur Nutzung und die Prüfung der Informationspakete sowie das Unterhalten von jeweils verwendeten Hard- und Softwarelösungen gehört natürlich zum Bereich des Managements im Open Archival Information System.
Der fünfte Teilbereich, der Bereich der Planung der Langzeitarchivierung im digitalen Archiv (Preservation Planning) befasst sich nicht nur mit der Sicherstellung des reibungslosen Informationszugangs in der Gegenwart, sondern ist vielmehr auf die Zukunft gerichtet. Es geht nämlich darum, Empfehlungen abzugeben, in welchen Zeitzyklen Updates vorgenommen werden müssen und in welchen Zyklen eine Migration der in einem Standardformat aufbewahrten elektronischen Aufzeichnungen in ein anderes neues Format vorgenommen werden müssen. Das heißt, eine ständige Überwachung im Bereich der Veränderung der Technologie gehört hier unabdingbar dazu. Aber auch der Blick auf den Benutzer und Veränderungen von Nutzungsgewohnheiten spielt hierbei eine Rolle. Preservation Planning umfasst dem zufolge die Erstellung von Vorlagen (Templates) für die Information Packages und die Entwicklung einer Migrationsstrategie im Archiv.
Der sechste und abschließende Bereich Zugriff (Access) befasst sich mit der Unterstützung der Benutzer beim Auffinden der entsprechenden elektronischen Informationen. Hier werden Anfragen entgegengenommen, Zugangsberechtigungen koordiniert und dann den jeweiligen Benutzergruppen die für sie nutzbaren Dissemination Information Packages, also Nutzungsinformationspakete, generiert und verteilt. Neben diesen fachlich ausgerichteten Aufgabenbereichen gehört natürlich auch ein Bereich der Verwaltung von OAIS als Gesamtsystem zum Betrieb und Unterhalt dazu, gewissermaßen die „Zentralabteilung“ des digitalen Archivs. Besondere Bedeutung hat dabei die Verwaltung der OAIS-Software, die nötig ist, um das Archiv überhaupt betreiben zu können. Dazu gehören der Aufbau eines funktionstüchtigen, aber auch geschützten Netzwerks, und die regelmäßige Überprüfung und Verbesserung der Sicherheit des

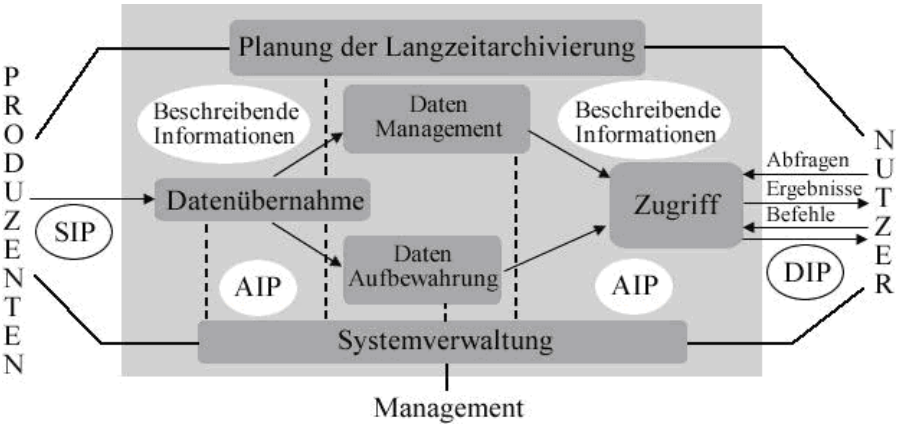
OAIS, um die in ihm enthaltenen Informationen vor unberechtigtem Zugang zu schützen.

Das OAIS setzt vollständig auf eine Migrationsstrategie als die derzeit von den Funktionen und der Technik her am besten beherrschbaren Strategie, selbst wenn es anderen Archivierungstechniken (z.B. Emulation) gegenüber offen ist. Migration wird im Sinne des OAIS in vier Bereiche systematisch zergliedert: erstens den Bereich des „Refreshment", des Wiederauffrischens mit dem Ziel, die Lesbarkeit der Datenträger zu sichern. Refreshment ist vor allen Dingen im Rahmen der AIPs, aber auch im Bereich der SIPs notwendig, damit überhaupt eine Übernahme möglich ist. Zum Refreshment tritt zweitens die „Replication", bei der regelmäßig der Kontext der verschiedenen Informationssysteme überprüft wird: Bestehende Verknüpfungen oder im Rahmen der Generierung von AIPs im Archiv hergestellte Verknüpfungen werden auf ihre Funktionstüchtigkeit und darauf überprüft, ob sie logisch schlüssig und verständlich sind. Ggf. ist drittens ein „Repackaging", also eine Art von digitaler Umbettung nötig, damit die bestehenden Verknüpfungen wieder funktionstüchtig sind oder ggf. neue Verknüpfungen erstellt werden (etwa dann, wenn vom Produzenten neue SIPs übernommen und zu AIPs umgeformt werden). Zum Schluss gehört auch die Transformation, d. h. die Übertragung auf neue, für einen bestimmten Zeitraum als tauglich erkannte Speichermedien, dazu. Hier wird im Rahmen des OAIS ein ganz zentraler Punkt angesprochen. Eine dauerhafte Lösung für die Langfristspeicherung, d.h. für die technische Sicherung der Zugänglichkeit wird auch in Zukunft nicht zu erwarten sein, sondern zur Archivierung digitaler Unterlagen wird es ab sofort gehören, immer mit den gegenwärtig zum technischen Standard gehörenden Informationsträgern leben zu müssen, die eine nur beschränkte Haltbarkeit haben und in Zukunft regelmäßig durch neue Formen von Informationsträgern ersetzt werden müssen. Es soll hier nur angedeutet werden, dass dieser Sachverhalt für eine Kostenplanung eines digitalen Archivs von entscheidender Bedeutung sein wird, weil nämlich neben eine Migration die der Sicherung des Zugangs dient, auch eine solche treten wird, die durch technische Innovationen im Hard- und Softwarebereich und eine weitere durch Veränderungen im Vorfeld des Archivs bedingt ist: Mit der Technik von gestern lassen sich digitale Objekte, die aus den gegenwärtigen Produktionssystemen stammen, nicht archivieren und langfristig zugänglich erhalten. Im Rahmen des OAIS verkennt man aber auch nicht, dass durch die skizzierte Migrationsstrategie Datenverluste möglich sind. Besonders im Bereich des Repackaging und der Transformation können diese Datenverluste auftreten. Man sieht aber im Augenblick noch keine realisierungsfähige technische Lösung, die diese Verluste vermeiden könnten.

4. Akzeptanz des OAIS-Modells

Das OAIS wird mittlerweile weltweit von Initiativen zur Langzeitarchivierung digitaler Ressourcen als Referenzmodell wahrgenommen und akzeptiert. Im Jahr 2002 wurde von der Niederländischen Nationalbibliothek in Den Haag der erste Prototyp eines digitalen Archivsystems (der gemeinsam mit IBM entwickelt wurde) in Dienst gestellt, das digitale Publikationen zugänglich halten soll. Dabei wurde das OAIS gezielt als Referenzmodell eingesetzt. Die Lösung ist großrechnerbasiert (IBM RS 6000S Winterhawk 2) und umfasst einen „Storage Server" mit 3,4 Tbyte Kapazität, sowie ein System redundanter Speicherung auf Optischen Medien (3x 1,3 Tbyte Kapazität) und Bandspeicherbibliotheken mit insgesamt 12 Tbyte Kapazität.
Das nationale Datenarchiv Großbritanniens (NDAD) hat seine Routinen und Prozeduren auf das OAIS umgestellt, und auch das australische Nationalarchiv orientiert sich im Rahmen des PANDORA-Projektes am OAIS.
Das amerikanische Nationalarchiv (NARA) hat die OAIS-Modellierung als Grundlage für die groß angelegte Ausschreibung zur Entwicklung des ehrgeizigen ERA-Systems (Electronic Records Archives) verwendet.
Standardisierungsaktivitäten für technische Metadaten zur Langzeiterhaltung und Kriterien für vertrauenswürdige digitale Archive verwenden Terminologie, Objekt- und Funktionsmodell von OAIS.

Anhang:



Grafik 7.1: Das Funktionsmodell des OAIS

SIP Submission Information Package = die digitalen Ressourcen, welche die aufbewahrenden Institutionen übernehmen.
AIP Archival Information Package = vom Langzeitarchiv mit Metadaten ergänzte digitale Medien. In dieser Form werden die digitalen Dokumente langfristig aufbewahrt.
DIP Dissemination Information Package = in dieser Form werden die digitalen Medien je nach rechtlichen Bedürfnissen generiert und zur Verfügung gestellt.

8 Vertrauenswürdigkeit von digitalen Langzeitarchiven

Abstract

Vertrauenswürdigkeit bildet ein zentrales Konzept beim Aufbau und bei der Bewertung digitaler Langzeitarchive. Neben organisatorischen Maßnahmen und Regelungen sind auch Sicherheitstechniken einsetzbar, die das Ziel haben, ebendiese Vertrauenswürdigkeit herzustellen.

8.1 Grundkonzepte der Sicherheit und Vertrauenswürdigkeit digitaler Objekte

Susanne Dobratz, Astrid Schoger und Niels Fromm

Bezogen auf das Ziel der digitalen Archivierung, die spätere Benutzbarkeit der Objekte zu erhalten und die Informationen zu sichern, finden im Laufe des Lebenszyklus eines digitalen Objektes verschiedene Methoden und Vorgehensweisen Anwendung. Diese werden heutzutage grob als Emulation und Migration bezeichnet. Durch die Anwendung dieser Methoden selbst, aber auch allein durch die Tatsache, dass die digitalen Objekte in einem Archivierungssystem verwaltet werden, sind sie spezielle Bedrohungen ausgesetzt.
Diese Bedrohungen können zum Beispiel sein, vgl. BSI, DRAMBORA, UNESCO, S. 31:

- Höhere Gewalt, wie etwa der Ausfall des IT-Systems, unzulässige Temperatur und Luftfeuchte, etc.;
- Organisatorische Mängel, wie Unerlaubte Ausübung von Rechten, Unzureichende Dokumentation von Archivzugriffen, Fehlerhafte Planung des Aufstellungsortes von Speicher- und Archivsystemen
- Menschliche Fehlhandlungen, wie Vertraulichkeits-/Integritätsverlust von Daten durch Fehlverhalten der IT-Benutzer, Verstoß gegen rechtliche Rahmenbedingungen beim Einsatz von Archivsystemen
- Technisches Versagen, wie Defekte Datenträger , Datenverlust bei erschöpftem Speichermedium, Verlust der Datenbankintegrität/-konsistenz , Ausfall oder Störung von Netzkomponenten, fehlerhafte Synchronisierung von Indexdaten bei der Archivierung, Veralten von Kryptoverfahren
- Vorsätzliche Handlungen, wie Manipulation an Daten oder Software, Anschlag, Unberechtigtes Kopieren der Datenträger, Sabotage, Unberechtigtes Überschreiben oder Löschen von Archivmedien

Ein Konzept zur Sicherung der Vertrauenswürdigkeit digitaler Objekte geht immer von der Annahme aus, dass die digitalen Objekte bestimmten Bedrohungen ausgesetzt sind und diese ein Risiko für die digitalen Objekte darstellen, dass es zu minimieren gilt, vgl. BSI 2005.

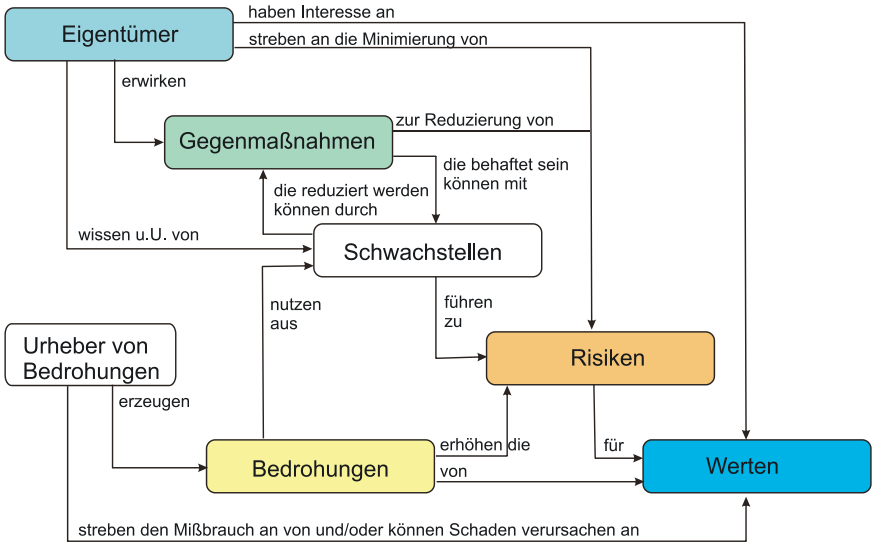


Abb. 8.1.1: Vertrauenswürdigkeitskonzeptgemäß den Common Criteria – Tafel 1

In der IT-Sicherheit, vgl. Steinmetz 2000 geht man davon aus, dass insbesondere folgende Eigenschaften eines digitalen Objektes bedroht sind und man zu deren Schutz entsprechende Maßnahmen ergreifen muss:

1. **Integrität**: bezeichnet den Aspekt, dass die digitalen Objekte unverändert vorliegen
2. **Authentizität**: bezieht sich auf den Aspekt der Nachweisbarkeit der Identität des Erstellers (Urhebers, Autors) und auf die Echtheit der digitalen Objekte
3. **Vertraulichkeit**: bezieht sich darauf, dass unberechtigten Dritten kein Zugang zu den digitalen Objekten gewährleistet wird.
4. **Verfügbarkeit**: bezieht sich auf den Aspekt der Zugänglichkeit zum digitalen Objekt unter Berücksichtigung der Zugriffsrechte
5. **Nichtabstreitbarkeit**: bezeichnet den Aspekt der Prüfung der Authentizität und Integrität digitaler Objekte durch berechtigte Dritte, sodass die Verbindlichkeit der Kommunikation gewährleistet wird, man nennt dies auch Authentifizierung.

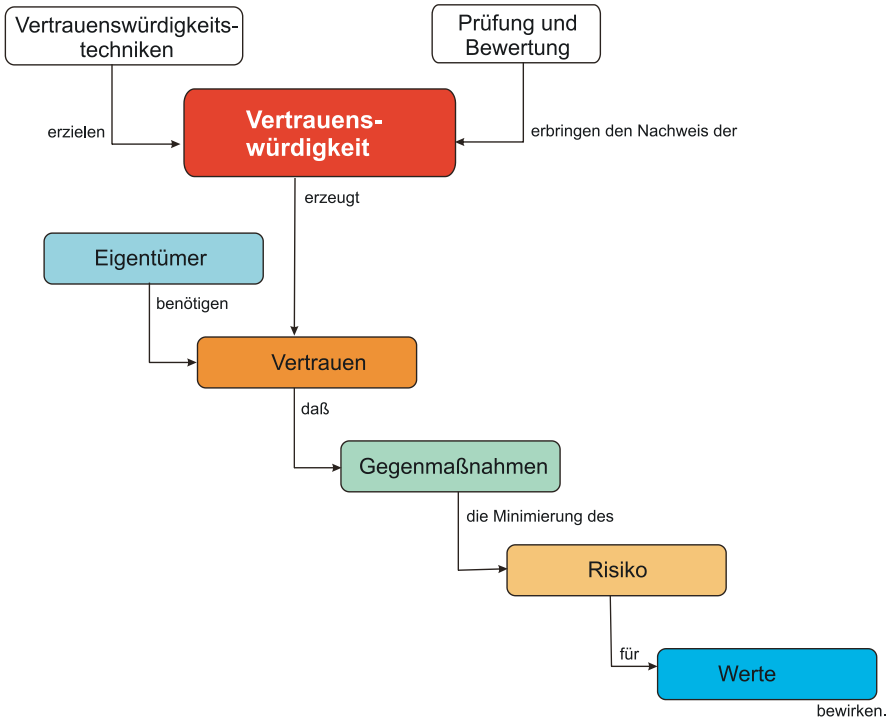


Abb. 8.1.2: Vertrauenswürdigkeitskonzeptgemäß den Common Criteria – Tafel 2

Digitale Langzeitarchive haben den Erhalt der Informationen über lange Zeiträume hinweg zum Ziel. Deshalb ergreifen sie sowohl organisatorische als auch technische Maßnahmen, um diesen Bedrohungen entgegenzuwirken.

Für die Sicherstellung der langfristigen Interpretierbarkeit, trotz der genannten Bedrohungen, ist die Integrität der archivierten digitalen Objekte von großer Bedeutung, da bei der Darstellung dieser Informationen schon wenige fehlerhafte Bits die gesamte Information unlesbar machen können. Zur Überprüfung der Unversehrheit digitaler Objekte, also deren Integrität, werden Hash- und Fingerprinting-Verfahren eingesetzt.
Für die Vertrauenswürdigkeit eines digitalen Langzeitarchivs stellen zudem die Authentizität und die Nichtabstreitbarkeit besonders wichtige Merkmale dar. Dies kann durch eine digitale Signatur der archivierten Objekte erreicht werden. Diese werden im nachfolgenden Kapitel dargestellt.

8.2 Praktische Sicherheitskonzepte

Dr. Siegfried Hackel, Tobias Schäfer, Dr. Wolf Zimmer

8.2.1 Hashverfahren und Fingerprinting

Ein wichtiger Bestandteil praktischer Sicherheitskonzepte zum Schutz der Integrität und Vertraulichkeit digitaler Daten sind Verschlüsselungsinfrastrukturen auf der Basis so genannter kryptographisch sicherer Hashfunktionen. Mit Hilfe kryptographisch sicherer Hashfunktionen werden eindeutige digitale „Fingerabdrücke" von Datenobjekten berechnet und zusammen mit den Objekten versandt oder gesichert. Anhand eines solchen digitalen „Fingerabdrucks" ist der Empfänger oder Nutzer der digitalen Objekte in der Lage, die Integrität eines solchen Objektes zur prüfen, bzw. unautorisierte Modifikationen zu entdecken.

Hashfunktionen werden in der Informatik seit langem eingesetzt, bspw. um im Datenbankumfeld schnelle Such- und Zugriffsverfahren zu realisieren. Eine Hashfunktion ist eine mathematisch oder anderweitig definierte Funktion, die ein Eingabedatum variabler Länge aus einem Urbildbereich (auch als „Universum" bezeichnet) auf ein (in der Regel kürzeres) Ausgabedatum fester Länge (den Hashwert, engl. auch message digest) in einem Bildbereich abbildet. Das Ziel ist, einen „Fingerabdruck" der Eingabe zu erzeugen, die eine Aussage darüber erlaubt, ob eine bestimmte Eingabe aller Wahrscheinlichkeit nach mit dem Original übereinstimmt.

Da der Bildbereich in der Regel sehr viel kleiner ist, als das abzubildende „Universum" können so genannte „Kollisionen" nicht ausgeschlossen werden. Eine Kollision wird beobachtet, wenn zwei unterschiedliche Datenobjekte des Universums auf den gleichen Hashwert abgebildet werden.

Für das Ziel, mit einer Hashfunktion einen Wert zu berechnen, der ein Datenobjekt eindeutig charakterisiert und damit die Überprüfung der Integrität von Daten ermöglicht, sind derartige Kollisionen natürlich alles andere als wünschenswert. Kryptographisch sichere Hashfunktionen H, die aus einem beliebig langen Wort M aus dem Universum von H einen Wert H(M), den Hashwert fester Länge erzeugen, sollen daher zwei wesentliche Eigenschaften aufweisen:

1. die Hashfunktion besitzt die Eigenschaften einer effizienten Ein-Weg-Funktion, d.h. für alle M aus dem Universum von H ist der Funktionswert h = H(M) effizient berechenbar und es gibt kein effizientes Verfah-

ren, um aus dem Hashwert h die Nachricht zu berechnen[1],

2. es ist - zumindest praktisch - unmöglich zu einem gegebenen Hashwert h = H(M) eine Nachricht M' zu finden, die zu dem gegebenen Hashwert passt (Urbildresistenz),
3. es ist - zumindest praktisch – unmöglich, zwei Nachrichten M und M' zu finden, die denselben Hashwert besitzen (Kollisionsresistenz).

Praktisch unmöglich bedeutet natürlich nicht praktisch ausgeschlossen, sondern bedeutet nicht mehr und nicht weniger, als dass es bspw. sehr schwierig ist, ein effizientes Verfahren zu finden, um zu einer gegebenen Nachricht M eine davon verschiedene Nachricht M' zu konstruieren, die denselben Hashwert liefert. Für digitale Objekte mit binären Zeichenvorräten Z = {0,1} lässt sich zeigen, dass für Hashfunktionen mit einem Wertbereich von 2^n verschiedenen Hashwerten, beim zufälligen Ausprobieren von $2^{n/2}$ Paaren von verschiedenen Urbildern M und M' die Wahrscheinlichkeit einer Kollision schon größer als 50% ist.

Beim heutigen Stand der Technik werden Hashfunktionen mit Hashwerten der Länge n = 160 Bit als hinreichend stark angesehen.[2] Denn, selbst eine Schwäche in der Kollisionsresistenz, wie bereits im Jahre 2005 angekündigt[3], besagt zunächst einmal lediglich, dass ein Angreifer zwei verschiedene Nachrichten erzeugen kann, die denselben Hashwert besitzen. Solange aber keine Schwäche der Urbildresistenz gefunden wird, dürfte es für einen Angreifer mit einem gegebenen Hashwert und passendem Urbild immer noch schwer sein, ein zweites, davon verschiedenes Urbild zu finden, das zu diesem Hashwert passt.

Kern kryptographischer Hashfunktionen sind Folgen gleichartiger Kompressionsfunktionen K, durch die eine Eingabe M blockweise zu einem Hashwert verarbeitet wird. Um Eingaben variabler Länge zu komprimieren, wendet man den Hashalgorithmus f iterierend an. Die Berechnung startet mit einem durch die Spezifikation des Hashalgorithmus festgelegten Initialwert $f(0):=I_0$. Anschließend gilt:

$f(i) := K(f(i-1),M_i)$ mit $M = M_1, \ldots, M_n$, $i = 1, \ldots, n$

1 Obwohl die Ein-Weg-Funktionen in der Kryptographie eine wichtige Rolle spielen, ist nicht bekannt, ob sie im streng mathematischen Sinne eigentlich existieren, ihre Existenz ist schwer zu beweisen. Man begnügt sich daher zumeist mit Kandidaten, für die man die Eigenschaft zwar nicht formal bewiesen hat, für die aber derzeit noch keine effizienten Verfahren zur Berechung der Umkehrfunktion bekannt sind.

2 Ein Rechner, der in der Lage ist, pro Sekunde den Hashwert zu einer Million Nachrichten zu berechnen, bräuchte 600.000 Jahre, um eine zweite Nachricht zu ermitteln, deren Hashwert mit einem vorgegebenen Hashwert der Länge 64 Bit übereinstimmt. Derselbe Rechner könnte allerdings in etwa einer Stunde irgendein Nachrichtenpaar mit gleichem Hashwert finden.

3 Schneier, B.: SHA-1 Broken, Feb. 2005, http://www.schneier.com

H(M):= f(n) = h ist der Hashwert von M

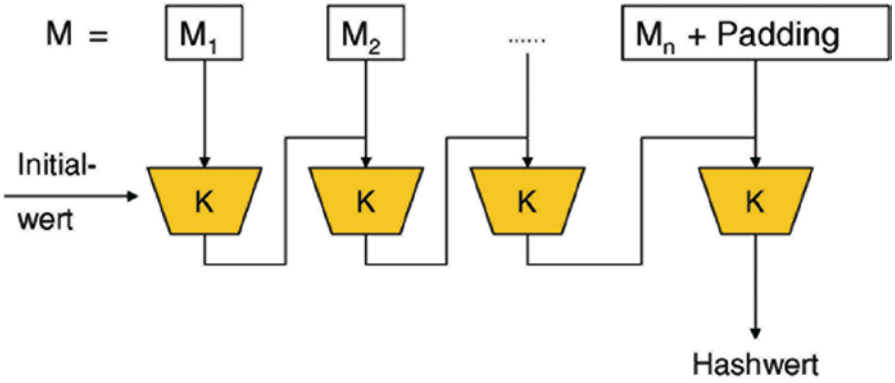


Abb. 8.2.1: Allgemeine Arbeitsweise von Hashfunktionen (nach C. Eckert[4])

Neben auf symmetrischen Blockchiffren, wie dem bereits 1981 durch das American National Standards Institute (ANSI) als Standard für den privaten Sektor anerkannten Data Encryption Standard (DES)[5], finden heute vor allem Hashfunktionen Verwendung, bei denen die Kompressionsfunktionen speziell für die Erzeugung von Hashwerten entwickelt wurden. Der bislang gebräuchlichste Algorithmus ist der Secure Hash Algorithm SHA-1 aus dem Jahre 1993.[6]
Der SHA-1 erzeugt Hashwerte von der Länge 160 Bits[7] und verwendet eine Blockgröße von 512 Bits, d. h. die Nachricht wird immer so aufgefüllt, dass die Länge ein Vielfaches von 512 Bit beträgt. Die Verarbeitung der 512-Bit Eingabeblöcke erfolgt sequentiell, für einen Block benötigt SHA-1 insgesamt 80 Verarbeitungsschritte.

8.2.2 Digitale Signatur

Elektronische Signaturen sind „Daten in elektronischer Form, die anderen elek-

4 Eckert, C.: IT-Sicherheit, Oldenburg Wissenschaftsverlag, 2001

5 vgl. bspw. Schneier, B.: Angewandte Kryptographie, Addison-Wesley Verl., 1996

6 vgl. bspw. Schneier, B.: ebenda

7 Da nicht ausgeschlossen werden kann, dass mit der Entwicklung der Rechentechnik künftig auch Hashwerte von der Länge 160 Bit nicht mehr ausreichend kollisions- und urbildresistent sind, wird heute für sicherheitstechnisch besonders sensible Bereiche bereits der Einsatz der Nachfolger SHA-256, SHA-384 und SHA-512 mit Bit-Längen von jeweils 256, 385 oder 512 Bits empfohlen.

tronischen Daten beigefügt oder logisch mit ihnen verknüpft sind und die zur Authentifizierung“ im elektronischen Rechts- und Geschäftsverkehr dienen. Ihre Aufgabe ist die Identifizierung des Urhebers der Daten, d.h. der Nachweis, dass die Daten tatsächlich vom Urheber herrühren (Echtheitsfunktion) und dies vom Empfänger der Daten auch geprüft werden kann (Verifikationsfunktion). Beides lässt sich nach dem heutigen Stand der Technik zuverlässig am ehesten auf der Grundlage kryptographischer Authentifizierungssysteme, bestehend aus sicheren Verschlüsselungsalgorithmen sowie dazu passenden und personifizierten Verschlüsselungs-Schlüsseln (den so genannten Signaturschlüsseln) realisieren.

Die Rechtswirkungen, die an diese Authentifizierung geknüpft werden, bestimmen sich aus dem Sicherheitsniveau, das bei ihrer Verwendung notwendig vorausgesetzt wird. Dementsprechend unterscheidet das im Jahre 2001 vom deutschen Gesetzgeber veröffentlichte „Gesetz über Rahmenbedingungen für elektronische Signaturen und zur Änderung weiterer Vorschriften“[8], kurz Signaturgesetz (SigG), vier Stufen elektronischer Signaturen:

- „Einfache elektronische Signaturen“ gem. § 2 Nr. 1 SigG,
- „Fortgeschrittene elektronische Signaturen“ gem. § 2 Nr. 2 SigG,
- „Qualifizierte elektronische Signaturen“ gem. § 2 Nr. 3 SigG,
- „Qualifizierte elektronische Signaturen“ mit Anbieter-Akkreditierung gem. § 15 Abs. 1 SigG.

Mit Ausnahme der einfachen elektronischen Signaturen, denen es an einer verlässlichen Sicherheitsvorgabe völlig fehlt, wird das mit der Anwendung elektronischer Signaturen angestrebte Sicherheitsniveau grundsätzlich an vier Elementen festgemacht (§ 2 Nr. 2 SigG). Elektronische Signaturen müssen demnach

- ausschließlich dem Signaturschlüssel-Inhaber zugeordnet sein,
- die Identifizierung des Signaturschlüssel-Inhabers ermöglichen,
- mit Mitteln erzeugt werden, die der Signaturschlüssel-Inhaber unter seiner alleinigen Kontrolle halten kann und
- mit den Daten, auf die sie sich beziehen, so verknüpft sein, dass eine nachträgliche Veränderung der Daten erkannt werden kann.

Europaweit als Ersatz für die handschriftliche Unterschrift akzeptiert werden jedoch lediglich qualifizierte elektronische Signaturen. Für sie wird zusätzlich gefordert (§ 2 Nr. 3 SigG), dass sie

- auf einem zum Zeitpunkt ihrer Erzeugung gültigen qualifizierten Zertifikat beruhen und
- mit einer sicheren Signaturerstellungseinheit erzeugt werden.

8 BGBl I 876; BT-Drs 14/4662 und 14/5324

Das Zertifikat übernimmt in diesem Fall die Authentizitätsfunktion, d. h. es bescheinigt die Identität der elektronisch unterschreibenden Person.[9] Sichere Signaturerstellungseinheiten sind nach dem Willen des Gesetzgebers Software- oder Hardwareeinheiten, die zur Speicherung und Anwendung des Signaturschlüssels dienen.[10]

Das Verfahren der digitalen Signatur basiert auf so genannten asymmetrischen kryptographischen Authentifizierungssystemen, bei denen jeder Teilnehmer ein kryptographisches Schlüsselpaar besitzt, bestehend aus einem geheimen privaten Schlüssel (private key, K_{priv}) und einem öffentlichen Schlüssel (public key, K_{pub}).

Eine wesentliche Eigenschaft solcher asymmetrischer Authentifizierungssysteme ist, dass es praktisch unmöglich ist, den privaten Schlüssel aus dem öffentlichen Schlüssel herzuleiten, der öffentliche Schlüssel wird durch Anwendung einer so genannten Einwegfunktion aus dem privaten Schlüssel berechnet. Der öffentliche Schlüssel kann daher in einem öffentlich zugänglichen Verzeichnis hinterlegt werden, ohne damit den privaten Schlüssel preiszugeben.

Der Urheber, respektive Absender elektronischer Daten „unterschreibt" nun seine Daten, indem er sie mit seinem geheimen, privaten Schlüssel verschlüsselt. Jeder, der die Daten empfängt, kann sie dann mit dem öffentlichen Schlüssel wieder entschlüsseln (s. Abb. 8.2.2).

9 Nach § 2 Nr. 6 SigG sind Zertifikate elektronische Bescheinigungen, mit denen Signaturschlüssel einer Person zugeordnet werden und die Identität einer Person bescheinigt wird. Für die Anwendung von Signaturverfahren von besonderer Bedeutung ist die Feststellung, dass „qualifizierte Zertifikate" nur auf natürliche Personen ausgestellt werden dürfen.

10 Das deutsche Signaturgesetz fordert, § 17 Abs. 1 SigG, dass sichere Signaturerstellungseinheiten vor unberechtigter Nutzung zu schützen sind. Nach § 15 Abs. 1 der Verordnung zur elektronischen Signatur (SigV) ist hierfür eine Identifikation „durch Besitz und Wissen oder durch Besitz und ein oder mehrere biometrische Merkmale" erforderlich. Da bislang keine Implementierungen biometrischer Verfahren bekannt sind, die die Anforderungen des Signaturgesetzes (vgl. Anlage 1 SigV) nachweislich erfüllen, werden für qualifizierte elektronische Signaturen in der Praxis immer Personal Identification Numbers (PIN) als Identifikationsdaten eingesetzt.

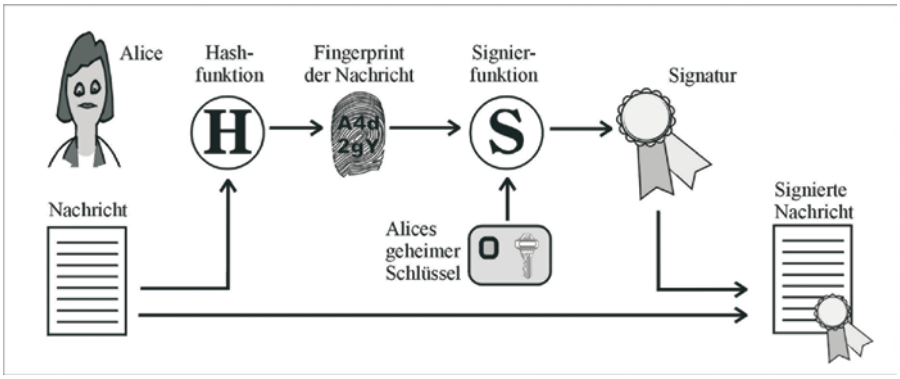


Abb. 8.2.2: Digitale Signatur

Unter der Voraussetzung, dass der öffentliche Schlüssel eindeutig und zuverlässig einer Person zugeordnet werden kann, bezeugt die Signatur folglich die Identität des Unterzeichners. Da die Signatur zudem das Ergebnis einer Verschlüsselungsoperation ist, sind die signierten Daten nachträglich auch nicht mehr veränderbar bzw. eine Änderung ist sofort erkennbar. Die Signatur kann auch nicht unautorisiert weiter verwendet werden, weil das Ergebnis der Verschlüsselungsoperation natürlich abhängig von den Daten ist. Geht man ferner davon aus, dass der private Signaturschlüssel nicht kompromittiert worden ist, kann der Absender der Daten die Urheberschaft auch nicht mehr zurückweisen, weil ausschließlich er selbst über den privaten Signaturschlüssel verfügt.
Technisch wäre natürlich eine Verschlüsselung der gesamten Daten (eines Dokuments oder einer Nachricht) viel zu aufwändig. Aus diesem Grunde wird aus den Daten eine eindeutige Prüfsumme, ein Hashwert (s. dazu auch Kap. 8.2.1) erzeugt, dieser verschlüsselt („unterschieben“) und den Originaldaten beigefügt. Der mit dem geheimen Schlüssel verschlüsselte Hashwert repräsentiert fortan die elektronische Signatur („Unterschrift“) der Originaldaten. Der Empfänger seinerseits bildet nach demselben Verfahren, d.h. mit demselben Hash-Algorithmus ebenfalls eine Prüfsumme aus den erhaltenen Daten und vergleicht sie mit der des Absenders. Sind die beiden Prüfsummen identisch, dann sind die Daten unverändert und stammen zuverlässig vom Inhaber des geheimen Schlüssels, denn nur er war in der Lage die Prüfsumme so zu verschlüsseln, dass sie mit dem zugehörigen öffentlichen Schlüssel auch entschlüsselt werden konnte.
Die Hinzufügung der Signaturdaten zu den Originaldaten kann grundsätzlich auf folgende Weise geschehen:

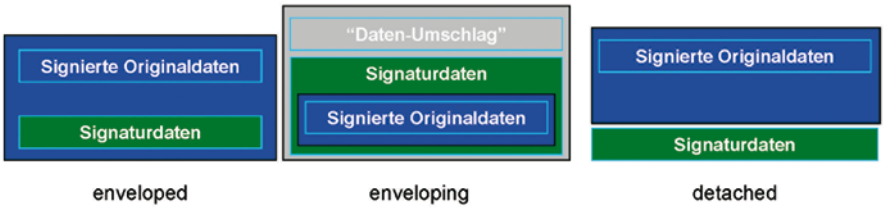


Abb. 8.2.3: Hinzufügung der Signaturdaten

- **Enveloped („eingebettet") :** die Signaturdaten sind als Element in den Originaldaten enthalten.
  Dieses Verfahren, auch als so genannte „Inbound-Signatur" bezeichnet, wird vor allem bei der Signatur von PDF-Dokumenten und PDF-Formularen bspw. im Projekt ArchiSafe der Physikalisch-Technischen Bundesanstalt benutzt (s. a. Abb. 8.2.4).[11] Dabei werden die binären Signaturdaten direkt in das PDF-Dokument eingebettet und gemeinsam mit den Originaldaten im PDF-Format angezeigt. Mit dem neuen Adobe® Reader® (Version 8) ist der Empfänger der signierten Daten darüber hinaus imstande, unmittelbar eine Überprüfung der Integrität der angezeigten und signierten Daten vorzunehmen.
  Eingebettete Signaturen werden ebenso bei der Signatur von XML-Daten[12] verwendet und sollen zudem nun auch für den neuen XDOMEA

11 http://www.archisafe.de

12 1999 bis 2002 wurde der W3C-Standard für das Signieren von XML-Dokumenten am Massachusetts Institute of Technology (MIT) entwickelt (XMLDSIG). Die XML Signatur Spezifikation (auch XMLDSig) definiert eine XML Syntax für digitale Signaturen.
In ihrer Funktion ähnelt sie dem PKCS#7 Standard, ist aber leichter zu erweitern und auf das Signieren von XML Dokumenten spezialisiert. Sie findet Einsatz in vielen weiterführenden Web-Standards wie etwa SOAP, SAML oder dem deutschen OSCI.
Mit XML Signaturen können Daten jeden Typs signiert werden. Dabei kann die XMLSignatur Bestandteil des XML Datenpakets sein (enveloped signature), die Daten können aber auch in die XML-Signatur selbst eingebettet sein (enveloping signature) oder mit einer URL adressiert werden (detached signature). Einer XML-Signatur ist immer mindestens eine Ressource zugeordnet, das heisst ein XML-Baum oder beliebige Binärdaten, auf die ein XML-Link verweist. Beim XML-Baum muss sichergestellt sein, dass es zu keinen Mehrdeutigkeiten kommt (zum Beispiel bezüglich der Reihenfolge der Attribute oder des verwendeten Zeichensatzes). Um dies erreichen zu können, ist eine so genannte Kanonisierung des Inhalts erforderlich. Dabei werden nach Maßgabe des Standards alle Elemente in der Reihenfolge ihres Auftretens aneinander gereiht und alle Attribute alphabetisch geordnet, so dass sich ein längerer UTF8-String ergibt (es gibt auch Methoden, die einen UTF16-String erzeugen). Aus

Standard 2.0[13] spezifiziert werden. Da die Signatur eine binäre Zahlenfolge ist, lässt sie sich jedoch nicht direkt in ein XML-Dokument einbetten. Man codiert daher die binären Werte im Base64-Format (RFC 1521), um aus ihnen ASCII-lesbare Zeichen zu gewinnen. Die erhaltene Zeichendarstellung der Signatur findet sich schliesslich als <SignatureValue> in der XML-Signatur wieder[14].

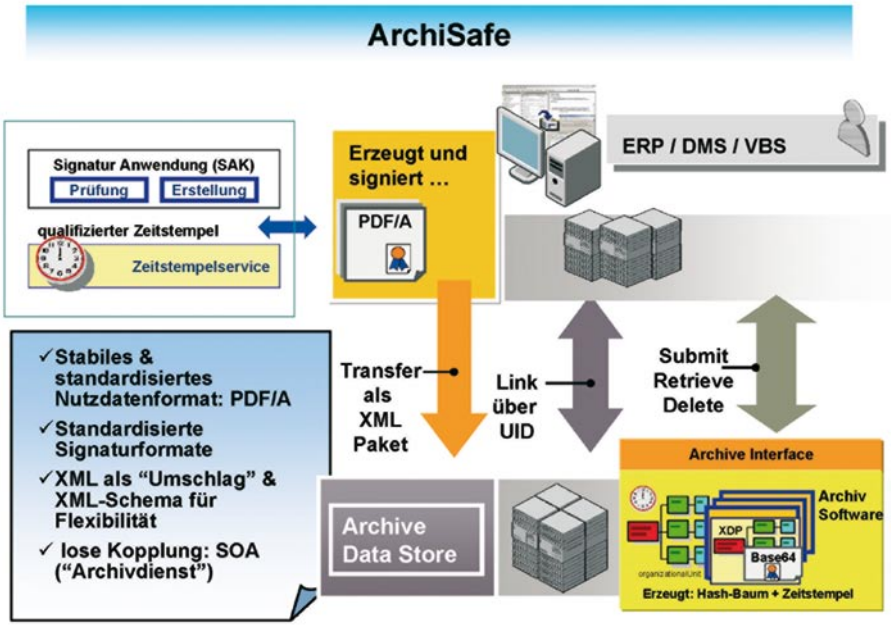


Abb. 8.2.4: ArchiSafe – Rechts- und revisionssichere Langspeicherung elektronischer Dokumente

- **Enveloping („umschließend"):** die Signaturdaten „umschließen" die Originaldaten. Diese Methode wird hauptsächlich für die Signatur von E-Mail Nachrichten oder reinen XML-Daten benutzt. Eine S/MIME Client-Anwendung, wie bspw. Microsoft Outlook, bettet in diesem Fall die Nachricht in einen signierten „Umschlag" ein.

diesem wird der eigentliche Hash-Wert gebildet beziehungsweise erzeugt man durch verschlüsseln den Signaturcode. So ist man wieder beim Standard-Verfahren für elektronische Signaturen (RFC 2437).

13 s. http://www.kbst.bund.de

14 Im Rahmen der Struktur eines XML-Dokuments lassen sich Subelemente explizit vom Signieren ausschliessen, so auch die Signatur selbst. Umgekehrt lassen sich beliebig viele Referenzen auflisten, die gemeinsam als Gesamtheit zu signieren sind.

- **Detached („getrennt"):** die Signaturdaten befinden sich außerhalb der Originaldaten in einer zusätzlichen, binären Signaturdatei. Diese Form, auch als „Outbound-Signatur" bezeichnet, wird standardmäßig für XML-Signaturen sowie die Signatur binärer Originaldaten eingesetzt. Ein separater Link in den Original-Daten oder zusätzlichen Beschreibungsdaten sorgt dann für die notwendige permanente Verknüpfung der Originaldaten mit den Signaturdaten.

Die Flexibilität der Hinzufügung von Signaturdaten zu Originaldaten basiert auf der als RFC 3852 – Cryptographic Message Syntax (CMS) im Juli 2004[15] durch die Internet Engineering Task Force (IETF) veröffentlichten Spezifikation sowie dem ursprünglich durch die RSA Laboratories veröffentlichten PKCS#7 (Public Key Cryptography Standard) Dokument in der Version 1.5. In beiden Dokumenten wird eine allgemeine Syntax beschrieben, nach der Daten durch kryptographische Maßnahmen wie digitale Signaturen oder Verschlüsselung geschützt, respektive Signaturdaten über das Internet ausgetauscht werden können. Die Syntax ist rekursiv, so dass Daten und Umschläge verschachtelt oder bereits chiffrierte Daten unterschrieben werden können. Die Syntax ermöglicht zudem, dass weitere Attribute wie z. B. Zeitstempel mit den Daten oder dem Nachrichteninhalt authentifiziert werden können und unterstützt eine Vielzahl von Architekturen für die Schlüsselverwaltung auf der Basis von elektronischen Zertifikaten.

15 Hously, R.: RFC 3852 – Cryptographic Message Syntax (CMS), Juli 2004, unter <http://www.ietf.org/rfc/rfc3852>

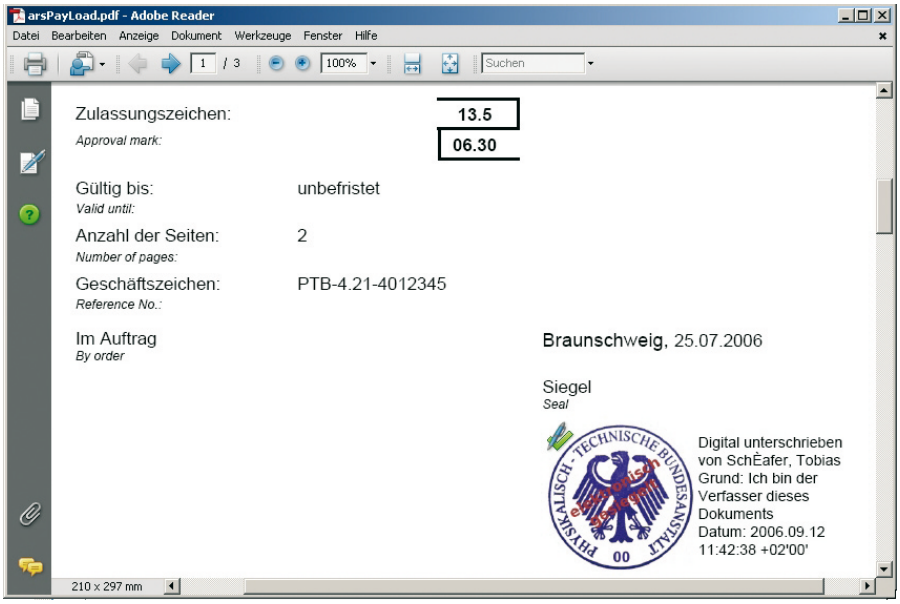


Abb. 8.2.5: Digitale PDF-Signatur

8.3 Evaluierung der Vertrauenswürdigkeit digitaler Archive

Susanne Dobratz und Astrid Schoger

8.3.1 Vertrauenswürdige Digitale Langzeitarchive

Die Anwendung von Methoden der IT-Sicherheit wie Hashfunktion und digitale Signatur kann bestimmte Risiken minimieren, insbesondere sie, die die Integrität, Authentizität und Vertraulichkeit digitaler Objekte betreffen. Das Problem einer breiten anwendung derartiger Technologien für die Langzeiterhaltung und vor allem für die Gewährleistung der Langzeitverfügbarkeit digitaler Objekte aus heutiger Sicht besteht vor allem darin, dass die langfristige Archivierung digitaler Signaturen technologisch nicht für beliebig große Datenmengen, komplexe Objekte und beliebig lange Zeiträume erprobt ist. Aktive Trustcenter, wie z.B. die Telesec GmbH[16] vergeben Zertifikate mit einer Gültigkeitsdauer von 3 Jahren. Danach müssen neue Zertifikate, die mit den alten verknüpft sind, ausgegeben und angewandt werden. Es handelt sich demnach um relativ kurzfristige Verfahren.

Dem gegenüber steht die Jahrhunderte lange Erfahrung der Archivare und Bibliothekare, die aus dem Blickpunkt der langfristigen Gewährleistung der Benutzbarkeit digitaler Objekte dem Einsatz digitaler Signaturen skeptisch gegenüberstehen.

Hier stehen die Aspekte der Verfügbarkeit und der Interpretierbarkeit digitaler Objekte in der Zukunft eine übergeordnete Rolle. Aus diesem Grunde konzentriert man sich darauf, organisatorische, wirtschaftlich-finanzielle Aspekte hervorzuheben und sich bei den technischen Aspekten auf die Methoden zu fokussieren, die die Anwendung von Normen und Standards bei der Abspeicherung der Objekte betreffen. Konkret sind dies die Aspekte des **Datenformats** und der **Metadaten** sowie der **Datenträger**, denen man eine besondere Bedeutung beimisst.

Daher haben verschiedene Organisationen und Initiativen mit der Formulierung von Anforderungen an vertrauenswürdige digitale Langzeitarchive begonnen. Diese Kriterien betreffen sowohl organisatorische als auch technische Rahmenbedingungen, die erfüllt werden müssen, um die Aufgabe der Erhaltung (der Interpretierbarkeit) digitaler Objekte gerecht werden zu können.

Dabei spielt die sogenannte Zielgruppe (engl. designated community) eine

16 Siehe http://ww.telesec.de

besondere Rolle, da z.B. die Interpretierbarkeit und die Nutzbarkeit digitaler Objekte auf die Vorkenntnisse, organisatorische und technische Benutzungsbedingungen und Nutzungsszenarien dieser Zielgruppe optimiert werden müssen. Die Anwendung konkreter Kriterien bzw. Anforderungen an das digitale Langzeitarchiv ist abhängig von der jeweiligen Zielgruppe.

Daher können allgemeingültige Anforderungen, wie sie die derzeit existierenden Kriterienkataloge darstellen, nur auf einem relativ abstrakten Niveau formuliert werden.

So hat die *nestor Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung* daher **Grundprinzipien bei der Herleitung und Anwendung der nestor-Kriterien** formuliert:

Abstraktion: Ziel des Kataloges ist es, Kriterien zu formulieren, die für ein breites Spektrum digitaler Langzeitarchive angewendet werden können und über längere Zeit Gültigkeit behalten sollen. Deshalb wird von relativ abstrakten Kriterien ausgegangen. Den Kriterien werden jeweils ausführliche Erläuterungen und konkrete Beispiele aus verschiedenen Bereichen mitgegeben. Die Beispiele entsprechen dem heutigen Stand der Technik und Organisation und sind unter Umständen nur im Kontext einer spezifischen Archivierungsaufgabe sinnvoll. Sie haben keinen Anspruch auf Vollständigkeit.

Dokumentation: Die Ziele, die Konzeption und Spezifikation sowie die Implementierung des digitalen Langzeitarchivs sind angemessen zu dokumentieren. Anhand der Dokumentation kann der Entwicklungsstand intern und extern bewertet werden. Eine frühzeitige Bewertung kann auch dazu dienen, Fehler durch eine ungeeignete Implementierung zu vermeiden. Insbesondere erlaubt es eine angemessene Dokumentation aller Stufen, die Schlüssigkeit eines digitales Langzeitarchiv umfassend zu bewerten. Auch alle Qualitäts- und Sicherheitsnormen fordern eine angemessene Dokumentation.

Transparenz: Transparenz wird realisiert durch die Veröffentlichung geeigneter Teile der Dokumentation. Transparenz nach außen gegenüber Nutzern und Partnern ermöglicht diesen, selbst den Grad an Vertrauenswürdigkeit festzustellen. Transparenz gegenüber Produzenten und Lieferanten bietet diesen die Möglichkeit zu bewerten, wem sie ihre digitalen Objekte anvertrauen. Die Transparenz nach innen dokumentiert gegenüber den Betreibern, den Trägern, dem Management sowie den Mitarbeitern die angemessene Qualität des digitalen Langzeitarchivs und sichert die Nachvollziehbarkeit der Maßnahmen. Bei denjenigen Teilen der Dokumentation, die für die breite Öffentlichkeit nicht geeignet sind (z.B. Firmengeheimnisse, Informationen mit Sicherheitsbezug), kann die Transparenz auf einen ausgewählten Kreis (z.B. zertifizierende Stelle) beschränkt werden. Durch das Prinzip der Transparenz wird Vertrauen aufge-

baut, da es die unmittelbare Bewertung der Qualität eines digitalen Langzeitarchivs durch Interessierte zulässt.
Angemessenheit: Das Prinzip der Angemessenheit berücksichtigt die Tatsache, dass keine absoluten Maßstäbe möglich sind, sondern dass sich die Bewertung immer an den Zielen und Aufgaben des jeweiligen digitalen Langzeitarchivs ausrichtet. Die Kriterien müssen im Kontext der jeweiligen Archivierungsaufgabe gesehen werden. Deshalb können ggf. einzelne Kriterien irrelevant sein. Auch der notwendige Erfüllungsgrad eines Kriteriums kann – je nach den Zielen und Aufgaben des digitalen Langzeitarchivs – unterschiedlich ausfallen.
Bewertbarkeit: Für die Vertrauenswürdigkeit existieren zum Teil - insbesondere unter Langzeitaspekten - keine objektiv bewertbaren (messbaren) Merkmale. In diesen Fällen ist man auf Indikatoren angewiesen, die den Grad der Vertrauenswürdigkeit repräsentieren. Transparenz macht auch die Indikatoren für eine Bewertung zugänglich.

8.3.2 Einige Definitionen

Die folgenden Begriffe sind im Zusammenhang mit vertrauenswürdigen digitalen Langzeitarchiven essentiell und orientieren sich am OAIS-Modell (siehe entsprechendes Kapitel im
Handbuch).

Digitales Objekt, Metadaten

Ein digitales Objekt ist eine logisch abgegrenzte Informationseinheit in der Form digitaler Daten. Daten sind maschinenlesbare und –bearbeitbare Repräsentationen von Information, in digitaler Form (eine Bitfolge, also eine Folge von Nullen und Einsen). Zur Nutzung der Informationen müssen die digitalen Daten interpretiert (dekodiert) werden.

Der Informationsbegriff umfasst hier jeden Typ von Wissen, der ausgetauscht werden kann; zum Beispiel aus inhaltlicher Sicht etwa Werke geistiger Schöpfung, Ergebnisse der Forschung und Entwicklung, Dokumentationen des politischen, sozialen und wirtschaftlichen Handelns.

Zu den Daten, die die Inhaltsinformation repräsentieren (Inhaltsdaten), können weitere Daten hinzukommen, die z.B. der Identifizierung, der Auffindbarkeit, der Rekonstruktion und Interpretation oder dem Nachweis der Integrität und Authentizität sowie der Kontrolle der Nutzungsrechte dienen (Metadaten). Metadaten können zu unterschiedlichen Zeiten im Lebenszyklus digitaler Objekte

entstehen (z.B. bei der Produktion, bei der Archivierung, bei der Bereitstellung für die Nutzung). Sie werden als Teile der logischen Einheit „digitales Objekt" aufgefasst und können sowohl getrennt als auch gemeinsam mit den Inhaltsdaten verwaltetet werden.

Digitales Langzeitarchiv, Vertrauenswürdigkeit

Unter einem digitalen Langzeitarchiv wird eine Organisation (bestehend aus Personen und technischen Systemen) verstanden, die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit zum Zwecke der Nutzung durch eine bestimmte Zielgruppe übernommen hat. Dabei bedeutet „Langzeit" über Veränderungen in der Technik (Soft- und Hardware) hinweg und auch unter Berücksichtigung möglicher Änderungen der Zielgruppe. Vertrauenswürdigkeit (engl. trustworthiness) wird als Eigenschaft eines Systems angesehen, gemäß seinen Zielen und Spezifikationen zu operieren (d.h. es tut genau das, was es zu tun vorgibt). Aus Sicht der IT-Sicherheit stellen Integrität, Authentizität, Vertraulichkeit und Verfügbarkeit Grundwerte dar. IT-Sicherheit ist somit ein wichtiger Baustein für vertrauenswürdige digitale Langzeitarchive.

8.3.3 Kriterienkataloge für vertrauenswürdige digitale Archive

Die Überprüfung und Bewertung der eingesetzten Maßnahmen zur Minimierung der Risiken, die den Langzeiterhalt der durch die digitalen Objekte repräsentierten Information bedrohen, erzeugt Vertrauenswürdigkeit. Diese kann anhand eines Kriterienkatalogs für Vertrauenswürdige digitaler Langzeitarchive geprüft und bewertet werden.
Dabei existieren internationale mehrer Ansätze.
Die Grundvoraussetzung für die Vertrauenswürdigkeit aller digitalen Langzeitarchive ist die, dass jedes nach seinen Zielen und Spezifikationen operiert. Diese sind durch die jeweilige Zielgruppe bestimmt.

Ein digitales Langzeitarchiv entsteht als komplexer Gesamtzusammenhang. Die Umsetzung der einzelnen Kriterien muss stets vor dem Hintergrund der Ziele des Gesamtsystems gesehen werden. Sowohl die Realisierung des digitalen Langzeitarchivs als Ganzes als auch die Erfüllung der einzelnen Kriterien läuft als Prozess in mehreren Stufen ab:

1. Konzeption

2. Planung und Spezifikation
3. Umsetzung und Implementation
4. Evaluierung

Im Zuge der ständigen Verbesserung sind diese Stufen nicht als starres Phasenmodell zu
betrachten sondern zu wiederholen.

Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)

Auf der Grundlage der Eigenschaften und Verantwortlichkeiten eines digitalen Langzeitarchivs, die bereits 2002 im Bericht der RLG/OCLC Working Group on Digital Archive Attributes als wesentlich für deren Vertrauenswürdigkeit aufgeführt wurden, hat die RLG-NARA Task Force on Digital Repository Certification 2006 als Entwurf in überarbeiteter Fassung 2007 eine Liste von Kriterien erarbeitet, die ein vertrauenswürdiges digitales Langzeitarchiv erfüllen müssen. Diese Liste dient der Orientierung, kann als Checkliste auch zur Selbstevaluierung und zum externen Audit eingesetzt werden, vgl. RLG-NARA 2007.

Nestor-Kriterienkatalog

nestor hat unter Berücksichtigung nationaler Ansätze und Arbeitsergebnisse wie des „DINI-Zertifikats für Dokumenten- und Publikationsserver“, vgl. DINI 2006, die Kriterien nationalen Rahmenbedingungen und den Bedürfnissen der deutschen Gedächtnisorganisationen angepasst und im Sommer 2006 als Entwurf zur öffentlichen Kommentierung veröffentlicht.
Überblick über die nestor-Kriterien:
Aus Gründen der Übersichtlichkeit wird im Folgenden der Term „digitales Langzeitarchiv“ mit „dLZA“ abgekürzt.

A. Organisatorischer Rahmen

1. Das dLZA hat seine Ziele definiert.

1.1 Das dLZA hat Kriterien für die Auswahl seiner digitalen Objekte entwickelt.

1.2 Das dLZA übernimmt die Verantwortung für den dauerhaften Erhalt der durch die digitalen Objekte repräsentierten Informationen.

1.3 Das dLZA hat seine Zielgruppe(n) definiert.

2. Das dLZA ermöglicht seinen Zielgruppe(n) eine angemessene Nutzung der durch die digitalen Objekte repräsentierten Informationen.

2.1 Das dLZA ermöglicht seinen Zielgruppe(n) den Zugang zu den durch die digitalen Objekte repräsentierten Informationen.

2.2 Das dLZA stellt die Interpretierbarkeit der digitalen Objekte durch seine Zielgruppe(n) sicher.

3. Gesetzliche und vertragliche Regelungen werden eingehalten.

3.1 Es bestehen rechtliche Regelungen zwischen Produzenten und dem digitalen Langzeitarchiv.

3.2 Das dLZA handelt bei der Archivierung auf der Basis rechtlicher Regelungen.

3.3 Das dLZA handelt bei der Nutzung auf der Basis rechtlicher Regelungen.

4. Die Organisationsform ist für das dLZA angemessen.

4.1 Die Finanzierung des digitalen Langzeitarchivs ist sichergestellt.

4.2 Es steht Personal mit angemessener Qualifikation in ausreichendem Umfang zur Verfügung.

4.3 Es bestehen angemessene Organisationsstrukturen für das dLZA.

4.4 Das dLZA betreibt eine langfristige Planung.

4.5 Die Fortführung der festgelegten Aufgaben ist auch über das Bestehen des digitalen Langzeitarchivs hinaus sichergestellt.

5. Es wird ein angemessenes Qualitätsmanagement durchgeführt.

5.1 Alle Prozesse und Verantwortlichkeiten sind definiert.

5.2 Das dLZA dokumentiert alle seine Elemente nach einem definierten Verfahren.

5.3 Das dLZA reagiert auf substantielle Veränderungen.

B. Umgang mit Objekten

6. Das dLZA stellt die Integrität der digitalen Objekte auf allen Stufen der Verarbeitung sicher.

6.1 Aufnahme (Ingest): Das dLZA sichert die Integrität der digitalen Objekte.

6.2 Archivablage (Archival Storage): Das dLZA sichert die Integrität der digitalen Objekte .

6.3 Nutzung (Access): Das dLZA sichert die Integrität der digitalen Objekte.

7. Das dLZA stellt die Authentizität der digitalen Objekte und Metadaten auf allen Stufen der Verarbeitung sicher.

7.1 Aufnahme (Ingest): Das dLZA sichert die Authentizität der digitalen Objekte.

7.2 Archivablage (Archival Storage): Das dLZA sichert die Authentizität der digitalen Objekte.

7.3 Nutzung (Access): Das dLZA sichert die Authentizität der digitalen Objekte.

8. Das dLZA betreibt eine langfristige Planung seiner technischen Langzeiterhaltungsmaßnahmen.

9. Das dLZA übernimmt digitale Objekte von den Produzenten nach definierten Vorgaben.

9.1 Das dLZA spezifiziert seine Übergabeobjekte (Submission Information Packages, SIPs).

9.2Das dLZA identifiziert, welche Eigenschaften der digitalen Objekte für den Erhalt von Informationen signifikant sind.

9.3Das dLZA erhält die physische Kontrolle über die digitalen Objekte, um Lang-zeit-archivierungsmaßnahmen durchführen zu können.

10. Die Archivierung digitaler Objekte erfolgt nach definierten Vorgaben.

10.1 Das dLZA definiert seine Archivobjekte (Archival Information Packages, AIPs).

10.2 Das dLZA sorgt für eine Transformation der Übergabeobjekte in Archivobjekte.

10.3 Das dLZA gewährleistet die Speicherung und Lesbarkeit der Archivobjekte.

10.4 Das dLZA setzt Strategien zum Langzeiterhalt für jedes Archivobjekt um.

11. Das dLZA ermöglicht die Nutzung der digitalen Objekte nach definierten Vorgaben.

11.1 Das dLZA definiert seine Nutzungsobjekte (Dissemination Information Packages, DIPs).

11.2 Das dLZA gewährleistet eine Transformation der Archivobjekte in Nutzungsobjekte.

12 Das Datenmanagement ist dazu geeignet, die notwendigen Funktionalitäten des digitalen Langzeitarchivs zu gewährleisten.

12.1 Das dLZA identifiziert seine Objekte und deren Beziehungen eindeutig und dauerhaft.

12.2 Das dLZA erhebt in ausreichendem Maße Metadaten für eine formale und inhaltliche Beschreibung und Identifizierung der digitalen Objekte.

12.3 Das dLZA erhebt in ausreichendem Maße Metadaten zur strukturellen Beschreibung der digitalen Objekte.

12.4 Das dLZA erhebt in ausreichendem Maße Metadaten, die die vom Archiv vorgenommenen Veränderungen an den digitalen Objekten verzeichnen.

12.5 Das dLZA erhebt in ausreichendem Maße Metadaten zur technischen Beschreibung der digitalen Objekte.

12.6 Das dLZA erhebt in ausreichendem Maße Metadaten, die die entsprechenden Nutzungsrechte und –bedingungen verzeichnen.

12.7 Die Zuordnung der Metadaten zu den Objekten ist zu jeder Zeit gegeben.

C. Infrastruktur und Sicherheit

13 Die ITInfrastruktur ist angemessen.

13.1 Die IT–Infrastruktur setzt die Forderungen aus dem Umgang mit Objekten um.

13.2 Die IT-Infrastruktur setzt die Sicherheitsanforderungen des IT-Sicherheitskonzeptes um.
14 Die Infrastruktur gewährleistet den Schutz des digitalen Langzeitarchivs und seiner digitalen Objekte.

10 gemeinsame Prinzipien und ISO

Die Kriterien für die Vertrauenswürdigkeit digitaler Langzeitarchive befinden sich zurzeit im Prozess internationaler Abstimmung und Standardisierung im Rahmen der ISO.
Wesentliche Vertreter des Themas Vertrauenswürdigkeit auf internationaler Ebene - Center of Research Libraries CRL, Digital Curation Centre DCC, Projekt Digital Preservation Europe DPE sowie nestor haben 10 gemeinsame Prinzipien herausgearbeitet, vgl. (CRL, DCC, DPE, nestor 2007), die den oben genannten Kriterienkatalogen und Audit Checklisten zu Grunde liegen. Diese stellen die Grundlage der weiteren Zusammenarbeit dar.
Ferner arbeitet eine internationale Arbeitsgruppe daran, die Kriterien für die Standardisierung im Rahmen der ISO vorzubereiten[17].

8.3.4 Wie wird evaluiert?

Orientierung, Selbstevaluierung, Audits

Die oben vorgestellten Kriterienkataloge und Checklisten dienen zurzeit zur Orientierung beim Aufbau digitaler Langzeitarchive und zur Selbstevaluierung sowie für externe Audits. Ein digitales Langzeitarchiv entsteht als komplexer Gesamtzusammenhang. Die Umsetzung der einzelnen Kriterien muss stets vor dem Hintergrund der Ziele des Gesamtsystems gesehen werden. Sowohl die Realisierung des digitales Langzeitarchivs als Ganzes als auch die Erfüllung der einzelnen Kriterien läuft als Prozess in mehreren Stufen ab: 1. Konzeption, 2. Planung und Spezifikation, 3. Umsetzung und Implementation, 4. Evaluierung. Im Zuge der ständigen Verbesserung sind diese Stufen nicht als starres Phasenmodell zu betrachten sondern zu wiederholen.

Digital Repository Audit Method based on Risk Assessment DRAMBORA

Im Rahmen des EU-Projektes Digitale Preservation Europe in Zusammenarbeit mit Digital Curation Centre wurde ein Tool zur Selbstevaluierung entwi-

17 In einer Birds Of Feather Gruppe unter Leitung von David Giaretta, vgl.

ckelt, das die Risikoanalyse als Methode einsetzt. Ausgehend von den Zielen eines digitalen Langzeitarchivs müssen zunächst die Werte und Aktivitäten spezifiziert, in einem weiteren Schritt dann die damit verbundenen Risiken identifiziert und bewertet werden.

Zertifizierung

Bevor ein international abgestimmtes Zertifizierungsverfahren für digitale Langzeitarchive entwickelt werden kann, muss zunächst ein internationaler Konsens über die Evaluierungskriterien gefunden werden. Ferner müssen aus den Erfahrungen mit der Anwendung der Kriterienkataloge und Evaluierungstools Bewertungsmaßstäbe für unterschiedliche Typen von digitalen Langzeitarchiven ausgearbeitet werden.

8.4 Literatur

RLG/OCLC Working Group on Digital Archive Attributes: Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report, 2002, http://www.rlg.org/en/pdfs/repositories.pdf

RLG-NARA Task Force on Digital Repository Certification, 2006: Audit Checklist for Certifying Digital Repositories, http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

RLG-NARA Task Force on Digital Repository Certification and CRL (2007): Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), http://www.crl.edu/PDF/trac.pdf

Deutsche Initiative für Netzwerkinformation/AG Elektronisches Publizieren (2006): DINI-Zertifikat für Dokumenten- und Publikationsserver, http://www.dini.de/documents/Zertifikat.pdf

„Kriterienkatalog vertrauenswürdige digitale Langzeitarchive Version 1 (Entwurf zur öffentlichen Kommentierung)“ herausgegeben von der nestor-Arbeitsgruppe Vertrauenswürdige Archive - Zertifizierung. (Frankfurt am Main: nestor-Materialien 8), 2006, http://nbn-resolving.de/urn:nbn:de:0008-2006060710

Gladney, Henry und Bennett, J. L. (2003): What Do We Mean by Authentic? What‘s the Real McCoy?, D-Lib Magazine (Band 9), Nr. 7/8. URL: DOI: 10.1045/july2003-gladney

Howard, John D. und Longstaff, Thomas A. (1998): A Common Language for Computer Security Incidents (Band SAND98-8667), SANDIA Reports. Auflage , Sandia National Laboratories, Albuquerque, New Mexico.

Oermann, Andrea und Dittmann, Jana (2008): Vertrauenswürdige und abgesicherte Langzeitarchivierung multimedialer Inhalte, nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland, Deutsche Nationalbibliothek, Frankfurt am Main.

CRL, DCC, DPE, nestor: Core Requirements for Digital Archives, (2007), http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92

DRAMBORA: Digital Repository Audit Method Based on Risk Assessment, (2007), http://www.repositoryaudit.eu/

Steinmetz, Ralf (2000): Multimedia-Technologie: Gundlagen, Komponenten und Systeme, 3. Auflage , Springer, Berlin, Heidelberg, New York.

BSI: Bundesamt für Sicherheit in der Informationstechnik (2005): Common Criteria V 2.3.

UNESCO (2003): Guidelines for the preservation of digital heritage, UNESCO, Paris.

9 Formate

Einleitung

Stefan E. Funk

Ein Computer-Programm muss die Daten, die es verwaltet, auf einen permanenten Datenspeicher (zum Beispiel eine CD oder eine Festplatte) ablegen, damit sie auch nach Ausschalten des Computers sicher verwahrt sind. Sie können so später erneut in den Rechner geladen werden. Um sicher zu stellen, dass ein geladenes Dokument exakt dem Dokument entspricht, welches zuvor gespeichert wurde, ist es erforderlich, dass das Programm die gesicherten Daten (sprich die Folge von Nullen und Einsen) exakt in der Weise interpretiert, wie es beim Speichern beabsichtigt war.

Um dies zu erreichen, müssen die Daten in einer Form vorliegen, die sowohl das speichernde als auch das ladende Programm gleichfalls „verstehen“ und interpretieren können. Ein Programm muss die Daten, die es verwaltet, in einem definierten *Dateiformat* speichern können. Dies bedeutet, alle zu speichernden Daten in eine genau definierte Ordnung zu bringen, um diese dann als eine

Folge von Bits zu speichern. Die Bits, mit denen beispielsweise der Titel eines Dokuments gespeichert ist, müssen später exakt von derselben Stelle und semantisch gesehen auch als Titel wieder in unser Programm geladen werden, wenn das Dokument seine ursprüngliche Bedeutung behalten soll. Somit muss das Programm das Format genau kennen, muss wissen, welche Bits des Bitstreams welche Bedeutung haben, um diese richtig zu interpretieren und verarbeiten zu können.

Ein *Format-Spezifikation* ist nun eine Beschreibung der Anordnung der Bits und somit eine Beschreibung, wie die Daten interpretiert werden müssen, um das ursprüngliche Dokument zu erhalten. Grob kann zwischen proprietären und offenen Dateiformaten unterschieden werden. Bei proprietären Dateiformaten ist die Spezifikation oft nicht bekannt und bei offenen Formaten ist die Spezifikation frei zugänglich. Aus einer Datei, dessen Format und Spezifikation bekannt ist, kann die gespeicherte Information auch ohne das vielleicht nicht mehr verfügbare lesende Programm extrahiert werden. Ist die Spezifikation nicht verfügbar, ist die Gefahr sehr groß, dass die enthaltenen Daten nicht mehr korrekt interpretiert werden können und so Informationen verloren gehen. Aus diesem Grund sind dokumentierte Spezifikationen und standardisierte Formate für die Langzeitarchivierung digitaler Daten sehr wichtig.

Als *Standard* bezeichnet man ein Formate, das sich entweder aus dokumentierten proprietären Formaten etabliert hat, weil es von sehr vielen Nutzern/Programmen aufgegriffen wurde, oder das speziell als Standard entwickelt wurde mit dem Ziel, den Datenaustausch zwischen Programmen oder Plattformen zu vereinfachen oder gar erst zu ermöglichen. Als Beispiele seien hier das Open Document Format (ODF) sowie Grafik-Formate wie TIFF (Tagged Image File Format), GIF (Graphics Interchange Format) und JPEG (Joint Photographic Experts Group) oder auch PDF (Portable Document Format) genannt.

9.1 Digitale Objekte

Stefan E. Funk

Die erste Frage, die im Zusammenhang mit der digitalen Langzeitarchivierung gestellt werden muss, ist sicherlich die nach den zu archivierenden Objekten. Welche Objekte möchte ich archivieren? Eine einfache Antwort lautet hier zunächst: digitale Objekte!

Eine Antwort auf die naheliegende Frage, was denn digitale Objekte eigentlich sind, gibt die Definition zum Begriff „digitales Objekt" aus dem OAIS[1]. Dieser Standard beschreibt ganz allgemein ein Archivsystem mit dessen benötigten Komponenten und deren Kommunikation untereinander, wie auch die Kommunikation vom und zum Nutzer. Ein digitales Objekt wird dort definiert als „An object composed of a set of bit sequences", also als ein aus einer Reihe von Bit-Sequenzen zusammengesetztes Objekt. Somit kann all das als ein digitales Objekt bezeichnet werden, das mit Hilfe eines Computers gespeichert und verarbeitet werden kann. Und dies entspricht tatsächlich der Menge der Materialien, die langzeitarchiviert werden sollen, vom einfachen Textdokument im .txt-Format über umfangreiche PDF-Dateien bis hin zu kompletten Betriebssystemen.

Ein digitales Objektes kann auf drei Ebenen beschrieben werden, siehe Abbildung:

- als physisches Objekt,
- als logisches Objekt und schließlich
- als konzeptuelles Objekt.

Ein digitales Objekt kann beispielsweise eine Datei in einem spezifischen Dateiformat sein, z.B. eine einzelne Grafik, ein Word-Dokument oder eine PDF-Datei. Als ein digitales Objekt können auch komplexere Objekte bezeichnet werden, wie Anwendungsprogramme wie Word oder Mozilla, eine komplette Webseite inkl. Text und Grafik, eine durchsuchbare Datenbank auf CD inklusive einer Suchoberfläche oder ein Betriebssystem wie Linux, Windows oder Mac OS .

Das physische Objekt - Daten auf einem Speichermedium

Als physisches Objekt sieht man die Menge der Zeichen an, die auf einem Informationsträger gespeichert sind. Die Art und Weise der physischen Be-

1 Open Archival Information System

schaffenheit dieser Zeichen kann aufgrund der unterschiedlichen Beschaffenheit des Trägers sehr unterschiedlich sein. Auf einer CD-ROM sind es die sogenannten „pits“ und „lands“ auf der Trägeroberfläche, bei magnetischen Datenträgern sind es Übergänge zwischen magnetisierten und nicht magnetisierten Teilchen. Auf der physischen Ebene haben die Bits keine weitere Bedeutung außer eben der, dass sie binär codierte Information enthalten, also entweder die „0“ oder die „1“. Auf dieser Ebene unterscheiden sich beispielsweise Bits, die zu einem Text gehören, in keiner Weise von Bits, die Teil eines Computerprogramms oder Teil einer Grafik sind.

Die Erhaltung dieses Bitstreams (auch Bitstreamerhaltung) ist der erste Schritt zur Konservierung des gesamten digitalen Objekts, er bildet sozusagen die Grundlage aller weiteren Erhaltungs-Strategien.

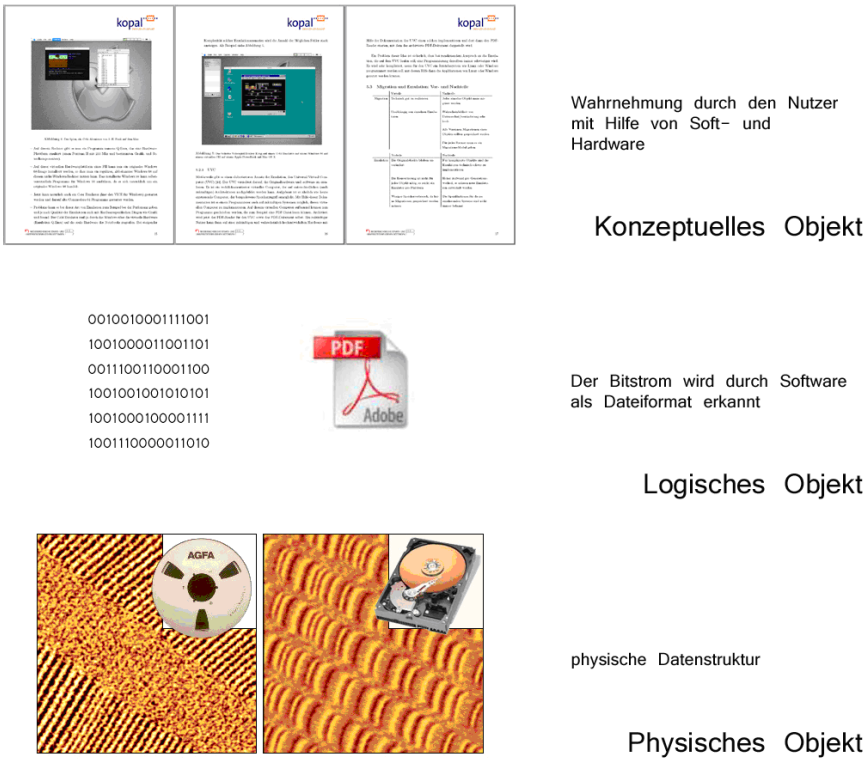
Das logische Objekt - Eine Folge von Bits als Einheit

Unter einem logischen Objekt versteht man eine Folge von Bits, die von einem Informationsträger gelesen und als eine Einheit angesehen werden kann. Diese können von einer entsprechenden Software als Format erkannt und verarbeitet werden. In dieser Ebene existiert das Objekt nicht nur als Bitstream, es hat bereits ein definiertes Format. Die Bitstreams sind auf dieser Ebene schon sehr viel spezieller als die Bits auf dem physischen Speichermedium. So müssen diese zunächst von dem Programm, das einen solchen Bitstream zum Beispiel als eine Textdatei erkennen soll, als eine solche identifizieren. Erst wenn der Bitstream als korrekte Textdatei erkannt worden ist, kann er vom Programm als Dateiformat interpretiert werden.

Will man diesen logischen Einheiten ihren Inhalt entlocken, muss das Format dieser Einheit genau bekannt sein. Ist ein Format nicht hinreichend bekannt oder existiert die zu dem Format gehörige Software nicht mehr, so wird die ursprüngliche Information des logischen Objektes sehr wahrscheinlich nicht mehr vollständig zu rekonstruieren sein. Um solche Verluste zu vermeiden, gibt es verschiedene Lösungsansätze, zwei davon sind Migration oder Emulation.

Das konzeptuelle Objekt - Das Objekt „zum Begreifen“

Das konzeptuelle Objekt beschreibt zu guter Letzt die gesamte Funktionalität, die dem Benutzer des digitalen Objekts mit Hilfe von dazu passender Soft- und Hardware zur Verfügung steht. Dies sind zunächst die Objekte, Zeichen und Töne, die der Mensch über seine Sinne wahrnimmt. Auch interaktive Dinge wie das



Die drei Ebenen eines digitalen Objekts

Spielen eines Computerspiels oder eine durchsuchbare Datenbank zählen dazu, denn die Funktion eines Computerspiels ist es, gespielt werden zu können. Ein weiteres Beispiel ist eine komplexe Textdatei, mit all ihren Editierungsmöglichkeiten, Tabellen und enthaltenen Bildern, die das verarbeitende Programm bietet.

Dieses konzeptuelle Objekt ist also die eigentliche, für den Betrachter bedeutungsvolle Einheit, sei es ein Buch, ein Musikstück, ein Film, ein Computerprogramm oder ein Videospiel. Diese Einheit ist es, die der Nachwelt erhalten bleiben soll und die es mit Hilfe der „Digital Preservation“ zu schützen gilt.

Die Erhaltung des konzeptuellen Objekts

Das Ziel eines Langzeitarchivs ist es also, das konzeptuelle Objekt zu archivieren und dem Nutzer auch in ferner Zukunft Zugriff auf dessen Inhalte zu gewähren. Die Darstellung bzw. Nutzung des digitalen Objekts soll so nahe wie

möglich den Originalzustand des Objekts zu Beginn der Archivierung wiederspiegeln. Dies ist nicht möglich, wenn sich Probleme bei der Archivierung auf den unteren Ebenen, der logischen und der physischen Ebene, ergeben. Gibt es eine unbeabsichtigte Veränderung des Bitstreams durch fehlerhafte Datenträger oder existiert eine bestimmte Software nicht mehr, die den Bitstream als Datei erkennt, ist auch eine Nutzung des Objekts auf konzeptueller Ebene nicht mehr möglich.

Literatur

- Reference Model for an Open Archival Information System (OAIS) <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf> (letzter Zugriff: 7. Juni 2006)
- Huth, Karsten, Andreas Lange: Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive (2004) <http://www.ichim.org/ichim04/contenu/PDF/2758_HuthLange.pdf> (letzter Zugriff: 7. Juni 2006)
- Thibodeau, K.: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In The State of Digital Preservation: An International Perspective. Washington D.C.: Council on Library and Information Resources. 4-31 (2001) <http://www.clir.org/PUBS/abstract/pub107abst.html> (letzter Zugriff: 12. Juli 2006)

9.2 Dateiformate

Stefan E. Funk

Dateiformate, in denen ein digitales Objekt vorliegt, um von Anwendungsprogrammen verarbeitet werden zu können, spielen bei der Archivierung eine große Rolle. Diese Formate sind mehr oder weniger klar spezifiziert, einige sind offene Standards und andere sind proprietäre Formate einzelner Firmen. Als Beispiele lassen sich hier Formate nennen wie PDF (Portable Document Format), XML (eXtensive Markup Language), HTML (HyperText Markup Language), DOC (Windows Document Format), verschiedene Bildformate wie TIF (Tagged Image Format) oder GIF (Graphic Interchange Format).

Formaterkennung
Will man solche Dokumente für die Nachwelt erhalten und den Zugriff auf deren Inhalte sichern, besteht die dringende Notwendigkeit, diese verschiedenen Formate zu kennen und zu erkennen. Es ist sehr wichtig zu wissen, welches Dateiformat ein digitales Dokument hat und ob das Format dieses Dokuments auch korrekt ist. Die Korrektheit dieser Daten stellt sicher, dass ein Dokument genutzt bzw. angezeigt und später im Sinne von Migration und Emulation bearbeitet werden kann. Bevor ein Objekt in ein Langzeitarchiv eingespielt wird, müssen spezifische Informationen über dieses Objekt vorhanden sein, sogenannte Metadaten, die genaue Aussagen darüber machen, welches Dateiformat in welcher Version vorliegt. Die Spezifikationen der unterschiedlichen Formate müssen hinreichend bekannt sein, um eine spätere Migration zu ermöglichen. Es reicht unter Umständen nicht aus, ein Dokument mit Hilfe eines Programmes anzeigen zu können, es sollte auch möglich sein, anhand der Spezifikationen ein Anzeige- oder Konvertierungsprogramm zu entwickeln.

Validation

Für die Langzeitarchivierung reicht es nicht aus zu wissen, dass eine Datei in einem bestimmten Format und in einer bestimmten Version dieses Formats vorliegt. Eine weitere wichtige Information ist die Korrektheit des Dokument im Sinne der Spezifikation dieses Formats. Nur so ist ein späteres Bearbeiten der Dokumente möglich, denn die Tools zur Konvertierung (oder Migration) bauen auf den Formatspezifiktionen auf. Habe ich beispielsweise ein Doku-

ment im PDF-Format der Version 1.2 vorliegen und prüfe nicht eingehend, ob dieses Format auch den Spezifikationen entspricht, könnte es sein, dass spätere Migrations- und Konvertierungs-Tools, die aus PDF 1.2 ein neueres Format (zum Beispiel PDF 1.6) erstellen sollen, das Dokument nicht richtig oder im schlimmsten Fall gar nicht verarbeiten können. Selbst wenn eine Datei korrekt dargestellt wird, ist noch nicht sichergestellt, dass sie auch der Formatspezifikation entspricht, da viele Anzeigeprogramme sehr fehlertolerant sind. Informationsverlust bis hin zum Verlust des gesamten Dokuments kann die Folge sein.

Metadaten

Zur Verwaltung von digitalen Objekten innerhalb eines Archivsystems werden Metadaten benötigt. Dies sind Daten über ein digitales Objekt. Zur Bestandserhaltung von digitalen Objekten werden zunächst technische Metadaten benötigt. Dies sind Daten wie Dateiformat und Version, Dateigröße, Dateiname, Checksumme zur Kontrolle der Integrität, MIME-Type, Erstellungsprogramm, Anzeigeprogramm, etc. Zur Dokumentation der Migrationsschritte dienen Provenance Metadaten. Diese beschreiben die Herkunft des Dokuments, beispielsweise die Art der Migration, den Zeitpunkt, die einzelnen durchgeführten Schritte und bei der Migration genutzte Programme. Deskriptive Metadaten beschreiben das Objekt inhaltlich, hierzu gehören unter anderem der Titel des Dokuments, der Name der Autoren, Abstract, Erscheinungsdatum und -Ort sowie Verlag. Rechtliche Metadaten schließlich beinhalten rechtliche Daten über das Dokument wie Eigentümer, Zugriffserlaubnis, etc.

Hilfsmittel

Es gibt Möglichkeiten, einige Metadaten maschinell zu erfassen. Die deskriptiven Metadaten zum Beispiel können aus den digitalen Katalogsystemen entnommen werden, sofern dafür geeignete Schnittstellen existieren. Die technischen Metadaten automatisch zu erfassen, ist in gewissen Grenzen ebenfalls möglich. Einige Programmier-Tools können technische Metadaten aus den digitalen Objekten extrahieren, zum Beispiel das Dateiformat und die Version desselben. Wie umfangreich die erhaltenen Metadaten sind, hängt von der Qualität des Tools ab. Im Einzelfall wird man solche Tools an die einzelnen Anforderungen anpassen müssen. Das Metadaten-Extraktions-Tool JHOVE[2] wird beispielsweise vom Projekt kopal[3] zur Erfassung von technischen Metadaten genutzt.

2 JSTOR/Harvard Object Validation Environment <http://hul.harvard.edu/jhove/index.html>

3 <http://kopal.langzeitarchivierung.de>

9.4 Formaterkennung und Validierung

Matthias Neubauer

Die Archivierung von digitalen Objekten steht und fällt mit der Erkennung und Validierung der verwendeten Dateiformate. Ohne die Information, wie die Nullen und Einsen des Bitstreams einer Datei zu interpretieren sind, ist der binäre Datenstrom schlicht unbrauchbar. Vergleichbar ist dies beispielsweise mit der Entzifferung alter Schriften und Sprachen, deren Syntax und Grammatik nicht mehr bekannt sind. Daher ist es für die digitale Langzeitarchivierung essentiell, die Dateien eines digitalen Objektes vor der Archivierung genauestens zu betrachten und zu kategorisieren. Dies beinhaltet vor allem zwei große Bereiche:

a) Die Formaterkennung

Zunächst muss das genaue Format ermittelt werden, in welchem die fragliche Datei vorliegt. Unterschiedliche Formate verwenden auch sehr unterschiedliche Identifizierungsmerkmale, was ein generell anwendbares Verfahren erschwert. Ein Merkmal, das zunächst nahe liegend erscheint, ist die so genannte Dateiendung oder File Extension. Dies bezeichnet den Teil des Dateinamens, welcher rechts neben dem letzten Vorkommen eines Punkt-Zeichens liegt (wie beispielsweise in „Datei.ext“). Dieses Merkmal ist jedoch meist nicht in einer Formatspezifikation festgelegt, sondern wird lediglich zur vereinfachten, oberflächlichen Erkennung und Eingruppierung von Dateien in Programmen und manchen Betriebssystemen genutzt. Vor allem aber kann die Dateiendung jederzeit frei geändert werden, was jedoch keinerlei Einfluss auf den Inhalt, und damit auf das eigentliche Format der Datei hat. Daher ist es nicht ratsam, sich bei der Formaterkennung allein auf die Dateiendung zu verlassen, sondern in jedem Fall noch weitere Erkennungsmerkmale zu überprüfen, sofern dies möglich ist. Einige Dateiformat-Spezifikationen definieren eine so genannte „Magic Number“. Dies ist ein Wert, welcher in einer Datei des entsprechenden Formats immer an einer in der Spezifikation bestimmten Stelle[4] der Binärdaten gesetzt sein muss. Anhand dieses Wertes kann zumindest sehr sicher angenommen werden, dass die fragliche Datei in einem dazu passenden Format vorliegt. Definiert ein Format keine „Magic Number“, kann meist nur durch den Versuch der Anwendung oder der Validierung der Datei des vermuteten Formats Klarheit darüber verschafft werden, ob die fragliche Datei tatsächlich in diesem

4 Eine bestimmte Stelle in einer Datei wird oft als „Offset“ bezeichnet und mit einem hexadezimalen Wert addressiert

Format abgespeichert wurde.

b) Die Validierung gegen eine Formatspezifikation

Die Validierung oder auch Gültigkeitsprüfung ist ein wichtiger und notwendiger Schritt vor der Archivierung von Dateien. Auch wenn das Format einer zu archivierenden Datei sicher bestimmt werden konnte, garantiert dies noch nicht, dass die fragliche Datei korrekt gemäß den Formatspezifikationen aufgebaut ist. Enthält die Datei Teile, die gegen die Spezifikation verstoßen, kann eine Verarbeitung oder Darstellung der Datei unmöglich werden. Besonders fragwürdig, speziell im Hinblick auf die digitale Langzeitarchivierung, sind dabei proprietäre und gegebenenfalls undokumentierte Abweichungen von einer Spezifikation, oder auch zu starke Fehlertoleranz eines Darstellungsprogrammes. Ein gutes Beispiel hierfür ist HTML, bei dem zwar syntaktische und grammatikalische Regeln definiert sind, die aktuellen Browser jedoch versuchen, fehlerhafte Stellen der Datei einfach dennoch darzustellen, oder individuell zu interpretieren. Wagt man nun einmal einen Blick in die „fernere" Zukunft - beim heutigen Technologiewandel etwa 20-30 Jahre - dann werden die proprietären Darstellungsprogramme wie beispielsweise die unterschiedlich interpretierenden Web-Browser Internet Explorer und Firefox wohl nicht mehr existieren. Der einzige Anhaltspunkt, den ein zukünftiges Bereitstellungssystem hat, ist also die Formatspezifikation der darzustellenden Datei. Wenn diese jedoch nicht valide zu den Spezifikationen vorliegt, ist es zu diesem Zeitpunkt wohl nahezu unmöglich, proprietäre und undokumentierte Abweichungen oder das Umgehen bzw. Korrigieren von fehlerhaften Stellen nachzuvollziehen. Daher sollte schon zum Zeitpunkt der ersten Archivierung sichergestellt sein, dass eine zu archivierende Datei vollkommen mit einer gegebenen Formatspezifikation in Übereinstimmung ist.

Sowohl für die aktuelle Bereitstellung der archivierten Dateien, als auch für spätere Migrations- und Emulationsszenarien ist demnach sowohl die Erkennung als auch die Validierung von Dateiformaten eine notwendige Voraussetzung. Ein Versäumnis dieser Aktionen kann einen erheblich höheren Arbeitsaufwand oder sogar einen vollkommenen Datenverlust zu einem späteren Zeitpunkt bedeuten.

9.5 File Format Registries

Andreas Aschenbrenner, Thomas Wollschläger

1. Zielsetzung und Stand der Dinge

Langzeitarchive für digitale Objekte benötigen aufgrund des ständigen Neuerscheinens und Veraltens von Dateiformaten aktuelle und inhaltlich präzise Informationen zu diesen Formaten. File Format Registries dienen dazu, den Nachweis und die Auffindung dieser Informationen in einer für Langzeitarchivierungsaktivitäten hinreichenden Präzision und Qualität zu gewährleisten. Da Aufbau und Pflege einer global gültigen File Format Registry für eine einzelne Institution so gut wie gar nicht zu leisten ist, müssen sinnvollerweise kooperativ erstellte und international abgestimmte Format Registries erstellt werden. Dies gewährleistet eine große Bandbreite, hohe Aktualität und kontrollierte Qualität solcher Unternehmungen.
File Format Registries können verschiedenen Zwecken dienen und dementsprechend unterschiedlich angelegt und folglich auch verschieden gut nachnutzbar sein. Hinter dem Aufbau solcher Registries stehen im Allgemeinen folgende Ziele:

- Formatidentifizierung
- Formatvalidierung
- Formatdeskription/-charakterisierung
- Formatlieferung/-ausgabe (zusammen mit einem Dokument)
- Formatumformung (z.B. Migration)
- Format-Risikomanagement (bei Wegfall von Formaten)

Für Langzeitarchivierungsvorhaben ist es zentral, nicht nur die Bewahrung, sondern auch den Zugriff auf Daten für künftige Generationen sicherzustellen. Es ist nötig, eine Registry anzulegen, die in seiner Zielsetzung alle sechs genannten Zwecke kombiniert. Viele bereits existierende oder anvisierte Registries genügen nur einigen dieser Ziele, meistens den ersten drei.

Beispielhaft für derzeit existierende File Format Registries können angeführt werden:
(I) die File Format Encyclopedia,
http://pipin.tmd.ns.ac.yu/extra/fileformat/
(II) FILExt,

http://filext.com/
(III) Library of Congress Digital Formats,
http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml
(IV) C.E. Codere's File Format site,
http://magicdb.org/stdfiles.html
(V) PRONOM,
http://www.nationalarchives.gov.uk/pronom/
(VI) das Global Digital Format Registry,
http://hul.harvard.edu/gdfr/
(VIIa) Representation Information Registry Repository,
http://registry.dcc.ac.uk/omar
(VIIb) DCC RI RegRep,
http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCRegRepV04
(VIII) FCLA Data Formats,
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

2. Bewertung von File Format Registries

Um zu beurteilen bzw. zu bewerten, ob sich spezielle File Format Registries für eine Referenzierung bzw. Einbindung in das eigene Archivsystem eignen, sollten sie sorgfältig analysiert werden. Sinnvoll können z.B. folgende Kriterien als Ausgangspunkt gewählt werden:

- Was ist der Inhalt der jeweiligen Registry? Wie umfassend ist sie aufgebaut?
- Ist der Inhalt vollständig im Hinblick auf die gewählte Archivierungsstrategie?
- Gibt es erkennbare Schwerpunkte?
- Wie werden Beschreibungen in die Registry aufgenommen? (Governance und Editorial Process)
- Ist die Registry langlebig? Welche Organisation und Finanzierung steckt dahinter?
- Wie kann auf die Registry zugegriffen werden?, Wie können ihre Inhalte in eine lokale Archivierungsumgebung eingebunden werden?

Künftig werden File Format Registries eine Reihe von Anforderungen adressieren müssen, die von den im Aufbau bzw. Betrieb befindlichen Langzeit-Archivsystemen gestellt werden. Dazu gehören u.a. folgende Komplexe:

I) Vertrauenswürdigkeit von Formaten

Welche Rolle spielt die qualitative Bewertung eines Formats für die technische Prozessierung? Braucht man beispielsweise unterschiedliche Migrationsroutinen für Formate unterschiedlicher Vertrauenswürdigkeit? Wie kann dann ein Kriterienkatalog für die Skalierung der confidence (Vertrauenswürdigkeit) eines Formats aussehen und entwickelt werden? Unter Umständen müssen hier noch weitere Erfahrungen mit Migrationen und Emulationen gemacht werden, um im Einzelfall zu einem Urteil zu kommen. Es sollte jedoch eine Art von standardisiertem Vokabular und Kriteriengebrauch erreicht werden und transparent sein.

II) Persistent Identifier

Wie können Persistent Identifier (dauerhafte und eindeutige Adressierungen) von File Formats sinnvoll generiert werden? So kann es bestimmte Vorteile haben, Verwandtschafts- und Abstammungsverhältnisse von File Formats bereits am Identifier ablesen zu können. Die Identifizierung durch „Magic Numbers" scheint zu diesem Zweck ebenso wenig praktikabel wie die anhand eventueller ISO-Nummern. Die vermutlich bessere Art der Identifizierung ist die anhand von Persistent Identifiers wie URN oder DOI.

III) ID-Mapping

Wie kann ein Mapping verschiedener Identifikationssysteme (Persistent Identifier, interne Identifier der Archivsysteme, ISO-Nummer, PRONOM ID, etc.) durch Web Services erreicht werden, um in Zukunft die Möglichkeit des Datenaustausches mit anderen File Format Registries zu ermöglichen?

IV) Integration spezieller Lösungen

Wie kann in die bisherigen nachnutzbaren Überlegungen anderer Institutionen die Möglichkeit integriert werden, spezifische Lösungen für den Datenaustausch bereit zu halten? Dies betrifft beispielsweise die Möglichkeit, lokale Sichten zu erzeugen, lokale Preservation Policies zuzulassen oder aber mit bestimmten Kontrollstatus von eingespielten Records (z.B. „imported", „approved", „deleted") zu arbeiten.

3. Bibliografie

- Abrams, Seaman: Towards a global digital format registry. 69th IFLA 2003. http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf
- Representation and Rendering Project: File Format Report. 2003. http://www.leeds.ac.uk/reprend/
- Lars Clausen: Handling file formats. May 2004. http://netarchive.dk/publikationer/FileFormats-2004.pdf

9.6 Tools

Matthias Neubauer

Wie bei jedem Vorhaben, das den Einsatz von Software beinhaltet, stellt sich auch bei der Langzeitarchivierung von digitalen Objekten die Frage nach den geeigneten Auswahlkriterien für die einzusetzenden Software-Tools.

Besonders im Bereich der Migrations- und Manipulationstools kann es von Vorteil sein, wenn neben dem eigentlichen Programm auch der dazugehörige Source-Code[5] der Software vorliegt. Auf diese Weise können die während der Ausführung des Programms durchgeführten Prozesse auch nach Jahren noch nachvollzogen werden, indem die genaue Abfolge der Aktionen im Source-Code verfolgt wird. Voraussetzung dafür ist natürlich, dass der Source-Code seinerseits ebenfalls langzeitarchiviert wird.

Nachfolgend werden nun einige Tool-Kategorien kurz vorgestellt, welche für die digitale Langzeitarchivierung relevant und hilfreich sein können.

a) Formaterkennung

Diese Kategorie bezeichnet Software, die zur Identifikation des Formats von Dateien eingesetzt wird. Die Ergebnisse, welche von diesen Tools geliefert werden, können sehr unterschiedlich sein, da es noch keine global gültige und einheitliche Format Registry gibt, auf die sich die Hersteller der Tools berufen können. Manche Tools nutzen jedoch schon die Identifier von Format Registry Prototypen wie PRONOM (beispielsweise „DROID“, eine Java Applikation der National Archives von Großbritannien, ebenfalls Urheber von PRONOM. Link: http://droid.sourceforge.net). Viele Tools werden als Ergebnis einen so genannten „MIME-Typ“ zurückliefern. Dies ist jedoch eine sehr grobe Kategorisierung von Formattypen und für die Langzeitarchivierung ungeeignet, da zu ungenau.

b) Metadatengewinnung

Da es für die Langzeitarchivierung, insbesondere für die Migrationsbemü-

5 Der Source- oder auch Quellcode eines Programmes ist die les- und kompilierbare, aber nicht ausführbare Form eines Programmes. Er offenbahrt die Funktionsweise der Software und kann je nach Lizenzierung frei erweiter- oder veränderbar sein (Open Source Software).

hungen, von großem Vorteil ist, möglichst viele Details über das verwendete Format und die Eigenschaften einer Datei zu kennen, spielen Tools zur Metadatengewinnung eine sehr große Rolle. Prinzipiell kann man nie genug über eine archivierte Datei wissen, jedoch kann es durchaus sinnvoll sein, extrahierte Metadaten einmal auf ihre Qualität zu überprüfen und gegebenenfalls für die Langzeitarchivierung nur indirekt relevante Daten herauszufiltern, um das Archivierungssystem nicht mit unnötigen Daten zu belasten. Beispiel für ein solches Tool ist „JHOVE" (das JSTOR/Harvard Object Validation Environment der Harvard University Library, Link: http://hul.harvard.edu/jhove/), mit dem sich auch Formaterkennung und Validierung durchführen lassen. Das Tool ist in Java geschrieben und lässt sich auch als Programmier-Bibliothek in eigene Anwendungen einbinden. Die generierten technischen Metadaten lassen sich sowohl in Standard-Textform, als auch in XML mit definiertem XML-Schema ausgeben.

c) Validierung

Validierungstools für Dateiformate stellen sicher, dass eine Datei, welche in einem fraglichen Format vorliegt, dessen Spezifikation auch vollkommen entspricht. Dies ist eine wichtige Voraussetzung für die Archivierung und die spätere Verwertung, Anwendung und Migration beziehungsweise Emulation dieser Datei. Das bereits erwähnte Tool „JHOVE" kann in der aktuellen Version 1.1e die ihm bekannten Dateiformate validieren; verlässliche Validatoren existieren aber nicht für alle Dateiformate. Weit verbreitet und gut nutzbar sind beispielsweise XML Validatoren, die auch in XML Editoren wie „Oxygen" (SyncRO Soft Ltd., Link: http://www.oxygenxml.com) oder „XMLSpy" (Altova GmbH, Link: http://www.altova.com/XMLSpy) integriert sein können.

d) Formatkorrektur

Auf dem Markt existiert eine mannigfaltige Auswahl an verschiedensten Korrekturprogrammen für fehlerbehaftete Dateien eines bestimmten Formats. Diese Tools versuchen selbstständig und automatisiert, Abweichungen gegenüber einer Formatspezifikation in einer Datei zu bereinigen, so dass diese beispielsweise von einem Validierungstool akzeptiert wird. Da diese Tools jedoch das ursprüngliche Originalobjekt verändern, ist hier besondere Vorsicht geboten!

Dies hat sowohl rechtliche als auch programmatische Aspekte, die die Frage aufwerfen, ab wann eine Korrektur eines Originalobjektes als Veränderung gilt, und ob diese für die Archivierung gewünscht ist. Korrekturtools sind üblicherweise mit Validierungstools gekoppelt, da diese für ein sinnvolles Korrekturverfahren unerlässlich sind. Beispiel für ein solches Tool ist „PDF/A Live!" (intarsys consulting GmbH, Link: http://www.intarsys.de/produkte/dokumententechnologien/pdf-a-live), welches zur Validierung und Korrektur von PDF/A konformen Dokumenten dient.

e) Konvertierungstools

Für Migrationsvorhaben sind Konvertierungstools, die eine Datei eines bestimmten Formats in ein mögliches Zielformat überführen, unerlässlich. Die Konvertierung sollte dabei idealerweise verlustfrei erfolgen, was jedoch in der Praxis leider nicht bei allen Formatkonvertierungen gewährleistet sein kann. Je nach Archivierungsstrategie kann es sinnvoll sein, proprietäre Dateiformate vor der Archivierung zunächst in ein Format mit offener Spezifikation zu konvertieren. Ein Beispiel hierfür wäre „Adobe Acrobat" (Adobe Systems GmbH, Link: http://www.adobe.com/de/products/acrobat/), welches viele Formate in PDF[6] überführen kann.

Für Langzeitarchivierungsvorhaben empfiehlt sich eine individuelle Kombination der verschiedenen Kategorien, welche für das jeweilige Archivierungsvorhaben geeignet ist. Idealerweise sind verschiedene Kategorien in einem einzigen Open Source Tool vereint, beispielsweise was Formaterkennung, -konvertierung und -validierung betrifft. Formatbezogene Tools sind immer von aktuellen Entwicklungen abhängig, da auf diesem Sektor ständige Bewegung durch immer neue Formatdefinitionen herrscht. Tools, wie beispielsweise „JHOVE", die ein frei erweiterbares Modulsystem bieten, können hier klar im Vorteil sein. Dennoch sollte man sich im Klaren darüber sein, dass die Archivierung von digitalen Objekten nicht mittels eines einzigen universellen Tools erledigt werden kann, sondern dass diese mit fortwährenden Entwicklungsarbeiten verbunden ist. Die in diesem Kapitel genannten Tools können nur Beispiele für eine sehr große Palette an verfügbaren Tools sein, die beinahe täglich wächst.

---

6 Portable Document Format, Adobe Systems GmbH, Link: <http://www.adobe.com/de/products/acrobat/adobepdf.html>

10 Standards und Standardisierungsbemühungen

10.1.1 Metadata Encoding and Transmission Standard: Das METS Abstract Model – Einführung und Nutzungsmöglichkeiten

Markus Enders

Einführung

Ausgehend von den Digitalisierungsaktivitäten der Bibliotheken Mitte der 90iger Jahre entstand die Notwendigkeit, die so entstandenen Dokumente umfassend zu beschreiben. Diese Beschreibung muß im Gegensatz zu den bis dahin üblichen Verfahrensweisen nicht nur einen Datensatz für das gesamte

Dokument beinhalten, sondern außerdem einzelne Dokumentbestandteile und ihre Abhängigkeiten zueinander beschreiben. Nur so lassen sich gewohnte Nutzungsmöglichkeiten eines Buches in die digitale Welt übertragen. Inhaltsverzeichnisse, Seitennummern sowie Verweise auf einzelne Bilder müssen durch ein solches Format zusammengehalten werden.
Zu diesem Zweck wurde im Rahmen des „Making Of Amerika" Projektes Ebind entworfen. Ebind selber war jedoch ausschließlich nur für Digitalisate von Büchern sinnvoll zu verwenden.
Um weitere Medientypen sowie unterschiedliche Metadatenformate einbinden zu können, haben sich Anforderungen an ein komplexes Objektformat ergeben. Dies setzt ein abstraktes Modell voraus mit Hilfe dessen sich Dokumente flexibel modellieren lassen und als Container Format verschiedene Standards eingebunden werden können. Ein solches abstraktes Modell bildet die Basis von METS und wird durch das METS-XML-Schema beschrieben. Daher wird METS derzeit auch fast ausschließlich als XML serialisiert in und Form von Dateien gespeichert. Als Container Format ist es in der Lage weitere XML-Schema (so genannte Extension Schemas) zu integrieren.

Das METS Abstract Model

Das METS „Abstract Model" beinhaltet alle Objekte innerhalb eines METS Dokuments und beschreibt deren Verhältnis zueinander. Zentraler Bestandteil eines METS-Dokuments ist eine Struktur. Diese Struktur kann eine logische oder physische Struktur des zu beschreibenden Dokumentes (bspw. eines Textes) abbilden. Das bedeutet, daß eine Struktur aus mindestens einer Struktureinheit (bspw. einer Monographie) besteht, die weitere Einheiten beinhalten kann. Somit läßt sich eine Struktur als Baum modellieren. In METS wird diese Struktur in der <structMap>-Sektion gespeichert. Jedes METS-Dokument kann mehrere Strukturen in separaten Sektionen beinhalten. So lassen sich bspw. logische und physische Strukturen voneinander trennen. In einer Struktur läßt sich das Inhaltsverzeichnis eines Werkes dokumentieren; in der anderen Struktur kann das Buch (mit Seiten als unterliegende Struktureinheiten) beschrieben werden. Das „Abstract Model" besitzt eine weitere Sektion – die <structLink> Sektion –, um Verweise zwischen unterschiedlichen Strukturen zu speichern.

Neben den Strukturen berücksichtigt das Modell auch Metadaten, wobei darunter nicht nur bibliographische Metadaten zu verstehen sind. Aus diesem Grund unterteilt das Modell die Metadaten in deskriptive Metadaten (in der Descriptive Metadata Section) und administrative Metadaten (in der Administrative Meta-

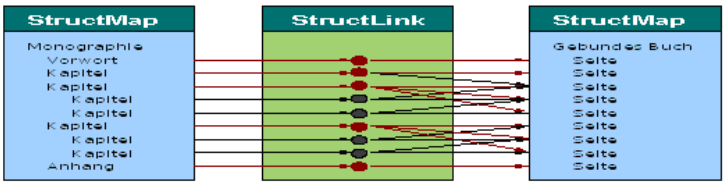


Abbildung 10.1.1.1: Verknüpfung von zwei Strukturen im Abstract-Model

Abbildung 10.1.1.2: Verknüpfung von zwei Strukturen im Abstract-Model

Abbildung 10.1.1.3: Administrative Metadata zu Dateien

Abbildung 10.1.1.4: Struktureinheit ist mit verschiedenen Dateien und Dateibereichen verknüpft

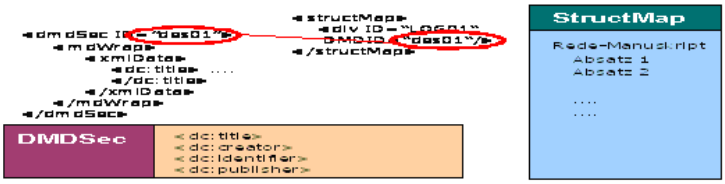


Abbildung 10.1.1.5: Unterschiedliche Sektionen mittels XML-IDs verknüpft

data Section). Während die deskriptiven Metadaten bibliographische Informationen enthalten, werden Informationen zu Rechteinhabern, Nutzungsrechte, technische Informationen zu einzelnen Dateien oder Langzeitarchivierungsmetadaten in den administrativen Metadaten gespeichert. Für beide Metadatentypen können beliebige Schema, so genannte „Extension Schema" genutzt werden, die in der jeweiligen Sektion gespeichert werden. Auf diese Weise lassen sich sowohl XML-Metadatenschema (MARC XML, MODS, Dublin Core simple) als auch Text-/ bzw. Binärdaten einbinden (bspw. PICA-Datensätze).

Neben den Struktureinheiten und ihren zugehörigen Metadaten spielen auch Dateien bzw. Streams eine wesentliche Rolle, da letztlich in ihnen die durch das METS-Dokument beschriebenen Inhalte gespeichert sind. Eine Datei kann bspw. den Volltext einen Buches, die Audioaufnahme einer Rede oder eine gescannte Buchseite als Image enthalten. Entsprechende Daten können in ein METS-Dokument eingebunden werden (bspw. BASE64 encoded in die METS-XML Datei eingefügt werden) oder aber mittels xlink referenziert werden. Ein METS-Dokument kann also als Container alle für ein Dokument notwendigen Dateien enthalten oder referenzieren, unabhängig davon, ob die Dateien lokal oder auf entfernten Servern vorhanden sind.
Grundsätzlich müssen alle für ein METS-Dokument relevanten Dateien innerhalb der File-Sektion aufgeführt werden. Innerhalb der File-Sektion können Gruppen (File-Groups) von Dateien gebildet werden, wobei die Abgrenzungskriterien zwischen einzelnen Gruppen nicht in METS fest definiert sind. Ja nach Modellierung lassen sich Dateien bspw. nach technischen Parametern (Auflösung oder Farbtiefe von Images), Anwendungszweck (Anzeige, Archivierung, Suche) oder sonstigen Eigenschaften (Durchlauf bestimmter Produktionsschritte) den einzelnen Gruppen zuordnen.
Das METS-Abstract-Model erlaubt das Speichern von administrativen Metadaten zu jeder Datei. Generelle, für jede Datei verfügbare technische Metadaten wie Dateigröße, Checksummen etc. lassen sich direkt in METS speichern. Für weiterführende Metadaten kann mit jeder Datei eine oder mehrere Administrative Metadatensektion(en) verknüpft werden, die bspw. Formatspezifische Metadaten enthalten (für Images könnten die Auflösungsinformationen, Informationen zur Farbtiefe etc. sein).

Dateien sind darüber hinaus mit Struktureinheiten verknüpft. Die Struktureinheit, die eine einzelne Buchseite repräsentiert, kann somit mit einer einzelnen Datei, die ein Image dieser Seite beinhaltet, verknüpft werden. Das „METS-Abstract-Model" stellt hierzu eine N:M Verknüpfung bereit. Das bedeutet, daß

eine Datei von mehreren Struktureinheiten (auch aus unterschiedlichen Strukturseksionen) aus verknüpft werden kann, genauso wie eine Struktureinheit mehrere Dateien verknüpfen kann. Im Ergebnis heißt das, daß der Struktureinheit vom Typ „Monographie“ sämtliche Imagedateien ein gescanntes Werken aus direkt unterstellt wird.
Für die Verknüpfung von Dateien sieht das „METS-Abstract-Model“ noch weitere Möglichkeiten vor. So lassen sich mehrere Verknüpfungen hinsichtlich ihrer Reihenfolge beim abspielen bzw. anzeigen bewerten. Dateien können entweder sequentiell angezeigt (Images eines digitalisierten Buches) als auch parallel abgespielt (Audio- und Videodateien gleichen Inhalts) werden. Darüber hinaus kann nicht nur auf Dateien, sondern auch in Dateiobjekte hinein verlinkt werden. Diese Verlinkungen sind u.a. dann sinnvoll, wenn Einheiten beschrieben werden, die aus technischen Gründen nicht aus der Datei herausgetrennt werden können. Das können bestimmte Teile eines Images sein (bspw. einzelne Textspalten) oder aber konkrete zeitliche Abschnitte einer Audioaufnahme. In der Praxis lassen sich so einzelne Zeitabschnitte eines Streams markieren und bspw. mit inhaltlich identischen Abschnitten eines Rede-Manuskriptes taggen. Das METS-Dokument würde über die Struktureinheit eine Verbindung zwischen den unterschiedlichen Dateien herstellen.

Das METS-Abstract-Model nutzt intensiv die Möglichkeit, einzelne Sektionen miteinander zu verknüpfen. Da METS überwiegend als XML realisiert ist, wird diese Verknüpfung über Identifier realisiert. Jede Sektionen verfügt über einen Identifier, der innerhalb des XML- Dokumentes eindeutig ist. Er dient als Ziel für die Verknüpfungen aus anderen Sektionen heraus. Aufgrund der XML-Serialisierung muß er weiteren Anforderungen genügen. Außerdem muß dieser Identifier mit den Regeln für XML-IDs verträglich sein. Ferner muß bei Verwendung von weiteren Extension Schemas darauf geachtet werden, daß die Eindeutigkeit der Identifier aus dem unterschiedlichen Schema nicht gefährdet wird, da diese üblicherweise alle im gleichen Namensraum existieren.

Wie deutlich geworden ist, stellt das METS-Abstract-Model sowie des XML-Serialisierung als METS-XML Schema lediglich ein grobes Modell da, welches auf den jeweiligen Anwendungsfall angepasst werden muß. Die Verwendung von Extension Schema sollte genauso dokumentiert werden wie die Nutzung optionaler Elemente und Attribute in METS. Dabei sollte vor allem auch die Transformation realer, im zu beschreibenden Dokument vorhandene Objekte in entsprechende METS-Objekte bzw. METS-Sektionen im Vordergrund stehen. Eine Strukturseksion kann bspw. lediglich logische Einheiten (bspw. das

Inhaltsverzeichnis eines Buches) umfassen als auch bestimmte physische Einheiten (bspw. einzelne Seiten) enthalten. Eine weitere Option wäre es, bestimmte Einheiten in eine zweite separate Struktur auszugliedern. Jede dieser Optionen mag für bestimmte Arten von Dokumenten sinnvoll sein.

Dokumentation

Damit ein METS-Dokument von unterschiedlichen Personen verstanden werden kann, ist es notwendig, neben den formalisierten METS-Schemas auch eine weitere Dokumentation der konkreten Implementierung von METS zu erstellen. Das METS-Profile-Schema bietet daher eine standarisierte Möglichkeit, eine solche Dokumentation zu erstellen, in dem sie eine Grobstrukturierung vorgibt und sicherstellt, daß alle wesentlichen Bereiche eines METS-Dokuments in der Dokumentation berücksichtigt werden.
Um ein solches Profil auf der offiziellen METS-Homepage veröffentlichen zu können, wird es durch Mitglieder des METS-Editorial-Board verifiziert. Nur verifizierte METS-Profile werden veröffentlicht und stehen auf der Homepage zur Nachnutzung bereit. Sie können von anderen Institutionen adaptiert und modifiziert werden und somit erheblich zur Reduktion der Entwicklungszeit einer eigenen METS-Implementierung beitragen.

Fazit

Aufgrund der hohen Flexibilität des METS Abstract Models wird METS in einer großen Zahl unterschiedlicher Implementierungen für sehr verschiedene Dokumententypen genutzt. Neben der ursprünglichen Anwendung, digitalisierte Büchern zu beschreiben, gibt es heute sowohl METS-Profile für Webseitenbeschreibungen (aus dem Bereich der Webseitenarchvierung) sowie Audio- und Videodaten. Während in den ersten Jahren METS überwiegend zum Beschreiben komplexer Dokumente genutzt wurde, um diese dann mittels XSLTs oder DMS-Systeme verwalten und anzeigen zu können, kommt heute METS gerade im Bereich der Langzeitarchivierung wachsende Bedeutung zu. METS ist heute für viele Bereiche, in denen komplexe Dokumente beschrieben werden müssen ein De-facto-Standard und kann sowohl im universitären als auch im kommerziellen Umfeld eine große Zahl an Implementierungen vorweisen. Ein großer Teil derer sind im METS-Implementation Registry auf der METS-Homepage (http://www.loc.gov/mets) nachgewiesen.

10.1.3 PREMIS

Olaf Brandt

PREMIS steht für „PREservation Metadata: Implementation Strategies“. Diese von der OCLC (Online Computer Library Center) und RLG (Research Library Group) im Jahre 2003 ins Leben gerufene Initiative betreibt die Entwicklung und Pflege des international anerkannten gleichnamigen PREMIS-Langzeitarchivierungsmetadatenstandards.

Die Mitglieder von PREMIS sind Akteure aus dem Umfeld von Gedächtnisorganisationen wie Archive, Bibliotheken und Museen, sowie der Privatwirtschaft. Diese befassen sich in internationalen Arbeitsgruppen mit Problemen der digitalen Langzeitarchivierung.

Das Hauptziel von PREMIS ist die Entwicklung von Empfehlungen, Vorschlägen und best-practices zur Implementierung von Langzeitarchivierungsmetadaten. Dazu gehört die Schaffung eines Kerns von Langzeitarchivierungsmetadaten mit größtmöglicher Anwendbarkeit innerhalb unterschiedlichster Langzeitarchivierungskontexte.

Die Arbeit von PREMIS baut auf den Ergebnissen der Preservation-Metadata-Working-Group auf. Diese Arbeitsgruppe wurde 2001 zur Entwicklung eines Rahmenkonzeptes für Langzeitarchivierungsmetadaten gebildet. Eine wichtige Grundlage für PREMIS ist das Referenzmodell des Open-Archival-Information-Systems (OAIS, ISO Standard 14721:2003). Dieses behandelt v.a. organisatorische und technische Fragen der digitalen Langzeitarchivierung.

Die Zielsetzung der so genannten Core Elements Group war bis Anfang 2005 die Entwicklung eines Kerns von Langzeitarchivierungsmetadaten, die Erstellung von Mappings und die Anbindung an andere Standards sowie der Aufbau eines Langzeitarchivierungsmetadatenlexikons. Die Ergebnisse dieser Gruppe sind in einem Abschlussbericht im Mai 2005 veröffentlicht worden.
Der Bericht beinhaltet das sogenannte PREMIS Data Dictionary 1.0, welches von einem ausführlichen Kommentar begleitet wird. Hierin sind der Kontext, das Datenmodell und die PREMIS-Grundannahmen aufgeführt. Zudem enthält der Bericht Erklärungen und Erläuterungen zu im Bericht erwähnten Themen, ein Glossar und erläuternde Beispiele. Das PREMIS-Data-Dictionary ist die Grundlage für die praktische Implementierung von Langzeitarchivierungs-

metadaten in digitalen Archiven.

Die zweite Arbeitsgruppe widmete sich den eher praktischen Fragen der realen Implementierung von Langzeitarchivierungsmetadaten. Untersucht wurden Fragen wie ‚Wie ist der Entwicklungsstand?' und ‚Was wird in welcher Weise implementiert?'. Darüber hinaus werden Themen über Datenhandling, eingesetzte Software und rechtliche Fragen erörtert. Erzielt wurden Empfehlungen zu best-practices auf Basis einer Reihe von Systemumgebungen. Die Ergebnisse flossen in einen im September 2004 veröffentlichten Untersuchungsbericht ein.

Implementierung

Aufbauend auf den Ergebnissen der Arbeitsgruppen stehen XML-Schemas zur Verfügung, welche in Langzeitarchivsysteme implementiert werden. Weiterhin sind die Schemas in Metadaten-Container-Formate (z.B. METS) integriert. Zu den nächsten Schritten zählen die maschinelle Erzeugung und Verarbeitung von PREMIS-Metadaten sowie die Integration in Workflows. Eine Liste von Institutionen, die PREMIS implementieren findet sich auf den PREMIS-Maintenance-Activity-Seiten der Library of Congress in den Vereinigten Staaten von Amerika. Eine rege Community in der Mailingliste der PREMIS-Implementors-Group diskutiert viele Fragen rund um die Implementierung von PREMIS und um unterschiedliche Themen der digitalen Langzeitarchivierung.
Die PREMIS-Maintenance-Activity übernimmt die weitere Koordination der Aktivitäten. Ein Teil davon, das PREMIS-Editorial-Committee, widmet sich der Verbreitung von PREMIS und der weiteren Pflege des Standards. Dazu gehören z.B. notwendige Anpassungen im Data-Dictionary oder den XML-Schemas. Diese Anpassungen werden gerade unter dem Eindruck der ersten praktischen Erfahrungen vorgenommen. Zur Verbreitung von PREMIS werden international unterschiedliche Veranstaltungen angeboten.
Den PREMIS-Aktivitäten wird im Kontext der Langzeitarchivierung übereinstimmend große Bedeutung im Bereich der Zusammenarbeit und des Datenaustausches beigemessen. Das schlägt sich auch in zwei internationalen Auszeichnungen nieder: für das Data-Dictionary wurde der PREMIS-Gruppe Ende 2005 den Digital-Preservation-Award der Digital-Preservation-Coalition und im August 2006 den Preservation-Publication-Award der Society of American Archivists verliehen.

Datenmodell

Das PREMIS-Datenmodell kennt einen vielseitigen Objektbegriff. Ein Objekt (Object) kann entweder eine Datei (File), ein Datenstrom (Bitstream) oder eine Repräsentation (Representation) sein. Ein Datenstrom ist dadurch gekennzeichnet, dass er sich nicht ohne zu ergänzende Daten oder einer Umformatierung in eine selbstständige Datei wandeln lässt. Eine Repräsentation ist eine Menge von Dateien, welche nur zusammenhängend eine sinnvolle und vollständige Darstellung einer intellektuellen Einheit (Intelectual Entity) liefern. Neben intellektuellen Einheiten und Objekten existieren im Datenmodell noch Rechte (Rights), Agenten (Agents) und Ereignisse (Events). Ereignisse und Rechte stehen in direkten Beziehungen zu Objekten und/oder Agenten. Zwischen Objekten können Beziehungen bestehen, die strukturelle Zusammengehörigkeit, Ableitungen oder Abhängigkeiten kennzeichnen.

Object Entity

Zu den Metadaten des Objekts gehören eindeutige Kennungen, Charakteristiken der Datei wie Größe und Format, Beschreibungen der Systemumgebungen (Software, Hardware), eine Auflistung der relevanten Eigenschaften der Objekte, sowie die Beziehungen zu Events und Rechteinformationen.

Event Entity

In der Ereignis-Entität können Aktionen, die in Verbindung mit Objekten oder Agenten stehen, dokumentiert werden. Dazu gibt es eindeutige Kennungen für Ereignisse und Aktionen, sowie Informationen über deren Resultate.

Agent-Entity

Ein Agent ist eine Person, eine Organisation oder Software, die Aktionen mit Objekten durchführt. Agenten werden durch eine eindeutige Kennung beschrieben.

Rights-Entity

Bei den Rechten werden Genehmigungen zur Durchführung von Aktionen von Agenten mit Objekten genau definiert.

Für PREMIS gibt es für jeden Entity-Typ ein eigenes XML-Schema, sodass eine modulare Einbindung in andere Schemas wie METS möglich ist.

Literatur:

Webseite der PREMIS Arbeitsgruppe: http://www.oclc.org/research/projects/pmwg/

Webseite der PREMIS Maintenance Activity: http://www.loc.gov/standards/premis/

Abschlußbericht der PREMIS Arbeitsgruppe inkl. Data Dictionary for Preservation Metadata: http://www.oclc.org/research/projects/pmwg/premis-final.pdf

PREMIS Survey Implementing Preservation Repositories for Digital Materials, Current Practice and Emerging Trends in the Cultural Heritage Community (survey report):

http://www.oclc.org/research/projects/pmwg/surveyreport.pdf

Digital Preservation Award 2005 der DPC:

http://www.dpconline.org/graphics/advocacy/press/award2005.html

Preservation Publication Award 2006 der Society of American Archivists:

http://www.archivists.org/recognition/dc2006-awards.asp#preservation

Preservation Metadata Working Group (PMWG 2002) Framework:

http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

10.1.4 LMER

Tobias Steinke

Die Langzeitarchivierungsmetadaten für elektronische Ressourcen (LMER) wurden von der Deutschen Bibliothek entwickelt. Das Objektmodell basiert auf dem „Preservation Metadata: Metadata Implementation Schema" der Nationalbibliothek von Neuseeland (2003).

Ziele von LMER sind:

- Ergänzung zu existierenden bibliographischen Metadaten, deshalb nur Beschreibung der technischen Informationen zu einem Objekt und der technischen Veränderungshistorie
- Praxisrelevante Beschränkung auf Angaben, die größtenteils automatisch generiert werden können
- Identifizierung der Kernelemente, die für alle Dateikategorien und jedes Dateiformat gültig sind, sowie ein flexibler Teil für spezifische Metadaten
- Abzubilden als XML-Schema
- Dateiformatidentifikation über Referenz zu einer zu schaffenden File-Format-Registry
- Modularer Aufbau zur Integration in Containerformate wie METS

Historie

LMER entstand 2003 aus dem Bedarf für technische Metadaten im Vorhaben LZA-RegBib. Die erste Version 1.0 wurde 2004 als Referenzbeschreibung und XML-Schema veröffentlicht. 2005 erschien eine überarbeitete Version 1.2, die auch Grundlage für die Verwendung im Projekt kopal ist. Die Version 1.2 führte eine starke Modularisierung und damit einhergehende Aufteilung in mehrere XML-Schemas ein, die eine bessere Einbindung in METS ermöglichte. Als Resultat entstand das METS-Profile-Universelles-Objektformat (UOF), das auf METS 1.4 und LMER 1.2 basiert.

Objektmodell

In LMER meint ein Objekt eine logische Einheit, die aus beliebig vielen Dateien bestehen kann. Es gibt einen Metadatenabschnitt zum Objekt und je einen Metadatenabschnitt zu jeder zugehörigen Datei. Zum Objekt einer jeder Datei

kann es Prozess-Abschnitte geben. Diese beschreiben die technische Veränderungshistorie, also vor allem die Anwendung der Langzeiterhaltungsstrategie Migration. Schließlich gibt es noch den Abschnitt Metadatenmodifikation, der Änderungen an den Metadaten selbst dokumentiert und sich auf alle anderen Abschnitte bezieht. Dabei wird davon ausgegangen, dass sich alle relevanten Metadatenabschnitte in derselben XML-Datei befinden.

Die vier möglichen Abschnittsarten LMER-Objekt, LMER-Datei, LMER-Prozess und LMER-Modifikation werden jeweils durch ein eigenes XML-Schema beschrieben. Dadurch kann jeder Abschnitt eigenständig in anderen XML-Schemas wie METS eingesetzt werden. Es gibt jedoch auch ein zusammenfassendes XML-Schema für LMER, das anders als die einzelnen Schemas Abhängigkeiten und Muss-Felder definiert.

LMER-Objekt

Die Metadaten zum Objekt stellen über einen Persistent Identifier den Bezug zu bibliographischen Metadaten her. Zugleich finden sich dort u.a. Informationen zur Objektversion und zur Anzahl der zugehörigen Dateien.

LMER-Datei

Zu jeder Datei werden die technischen Informationen erfasst, wie sie auch von einem Dateisystem angezeigt werden (Name, Pfad, Größe, Erstellungsdatum), aber auch eine Referenz zu exakten Formatbestimmung. Zudem wird jede Datei einer Kategorie zugeordnet (Bild, Video, Audio, etc.), die insbesondere für die spezifischen Metadaten relevant ist. Denn in einem speziellen Platzhalterelement des Datei-Abschnitts können dank des flexiblen Mechanismus von XML-Schemata beliebige XML-Metadaten zur spezifischen Bestimmung bestimmter Dateicharakteristiken hinterlegt werden. Ein Beispiel dafür ist die Ausgabe des Dateianalysewerkzeugs JHOVE.

LMER-Prozess

Die Metadaten in einem Prozess-Abschnitt beschreiben die Schritte und Resultate von technischen Veränderungen und Konvertierungen (Migrationen) an einem Objekt oder einzelnen Dateien eines Objekts. Gehört ein Prozess-Abschnitt zu einem Objekt, so bezeichnet er auch die Versionsnummer und die

Kennung des Objekts, von dem die vorliegende Version abgeleitet wurde.

LMER-Modifikation

Die LMER-Daten werden in der Regel in einer oder mehreren XML-Dateien gespeichert. Veränderungen (Ergänzungen oder Korrekturen) der XML-Daten darin können im Modifikationsabschnitt aufgeführt werden.

Literatur

Referenzbeschreibung zu LMER 1.2:

http://nbn-resolving.de/?urn=urn:nbn:de:1111-2005041102

Referenzbeschreibung zum Universellen Objektformat (UOF):

http://kopal.langzeitarchivierung.de/downloads/kopal_Universelles_Objektformat.pdf

10.1.5 MIX

Tobias Steinke

MIX steht für „NISO Metadata for Images in XML" und ist ein XML-Schema für technische Metadaten zur Verwaltung von digitalen Bildersammlungen. Die Metadatenelemente dieses XML-Schemas werden durch den Standard ANSI/ NISO Z39.87-2006 („Technical Metadata for Digital Still Images") beschrieben. MIX wurde von der Library of Congress und dem MARC Standards Office entwickelt. Neben allgemeinen Informationen zu einer Datei werden insbesondere komplexe Informationen zu Bildeigenschaften wie Farbinformationen aufgenommen, sowie detailierte Beschreibungen der technischen Werte der Erzeugungsgeräte wie Scanner oder Digitalkamera. Zusätzlich kann eine Veränderungshistorie in den Metadaten aufgeführt werden, wobei dies ausdrücklich als einfacher Ersatz für Institutionen gedacht ist, welche keine eigenen Langzeitarchivierungsmetadaten wie PREMIS nutzen. Es gibt keine Strukturinformationen in MIX, denn hierfür wird das ebenfalls von der Library of Congress stammende METS vorgesehen. Die aktuelle Version von MIX ist 1.0 von 2006. Ein öffentlicher Entwurf für MIX 2.0 liegt vor.

Offizielle Webseite: http://www.loc.gov/standards/mix/

11 Hardware

11.1 Hardware-Environment

Dagmar Ullrich

Abstract

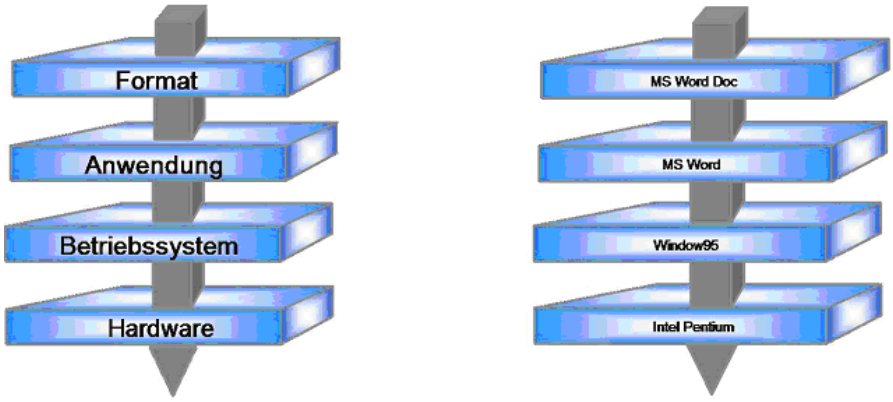
Digitale Datenobjekte benötigen eine Interpretationsumgebung, um ihren Inhalt für Menschen zugänglich zu machen. Diese Umgebung kann in unterschiedliche Schichten gegliedert werden, deren unterste die Hardware-Umgebung bildet. Diese Einteilung wird anhand eines Schichtenmodells, dem „Preservation Layer Model" veranschaulicht. Die Hardware-Umgebung umfasst nicht nur eine geeignete Rechnerarchitektur zur Darstellung der Inhalte, sondern auch eine funktionsfähige Speicherumgebung für den physischen Erhalt und die Bereitstellung des digitalen Datenobjektes.

Gliederung

Interpretationsumgebung digitaler Objekte und „Preservation Layer Model"
Speicherung und Bereitstellung des digitalen Objekts

Interpretationsumgebung digitaler Objekte und „Preservation Layer Model“

Um ein digitales Datenobjekt lesbar zu halten, muss eine entsprechende Interpretationsumgebung verfügbar sein. Diese umfasst Hardware, Betriebssystem und Anwendungssoftware. Um z.B. eine Word-Datei anzuzeigen wird eine passende Version von MS-Word benötigt. Für die Installation der Anwendungssoftware muss ein geeignetes Betriebssystem verfügbar sein, das seinerseits auf eine entsprechende Rechnerarchitektur angewiesen ist. In der Regel gibt es mehrere mögliche Kombinationen. Die Lesbarkeit digitaler Daten ist nur so lange sichergestellt, wie mindestens eine solche gültige Kombination einsatzfähig ist. Dieser Zusammenhang wird im Konzept des „Preservation Layer Models“ herausgearbeitet. Die nachstehende Grafik veranschaulicht dieses Konzept.[1]



Preservation Layer Model — Beispiel: View Path

Eine funktionsfähige Kombination der verschiedenen Ebenen wird als gültiger „View Path“ eines digitalen Datenobjektes bezeichnet und kann dem entsprechenden Objekt zugeordnet werden. Das Preservation Layer Model wurde an der Nationalbibliothek der Niederlande gemeinsam mit IBM entwickelt, um rechtzeitig zu erkennen, wann ein Datenobjekt Gefahr läuft, ohne gültigen View Path und damit nicht mehr lesbar zu sein. Zeichnet sich der Wegfall einer Komponente ab, lässt sich automatisch feststellen, welche View Paths und

1 Eine ausführliche Beschreibung des Preservation Layer Models findet sich in: Van Diessen, Raymond J. (2002): *preservation requirements in a deposit system.* Amsterdam: IBM Netherlands. S. 7-15. http://www-05.ibm.com/nl/dias/resource/preservation.pdf [2007, 20. August]

somit welche Datenobjekte betroffen sind. Auf dieser Grundlage kann dann entweder eine Emulationsstrategie entwickelt oder eine Migration betroffener Datenobjekte durchgeführt werden. Im Falle einer Formatmigration werden alle darunter liegenden Ebenen automatisch mit aktualisiert. Die Hard- und Softwareumgebung des alten Formats wird nicht mehr benötigt. Will man jedoch das Originalformat erhalten, müssen auch Betriebssystem und Rechnerarchitektur als Laufzeitumgebung der Interpretationssoftware vorhanden sein. Nicht immer hat man die Wahl zwischen diesen beiden Möglichkeiten. Es gibt eine Reihe digitaler Objekte, die sich nicht oder nur mit unverhältnismäßig hohem Aufwand in ein aktuelles Format migrieren lassen. Hierzu gehören vor allem solche Objekte, die selbst ausführbare Software enthalten, z.B. Informationsdatenbanken oder Computerspiele. Hier ist die Verfügbarkeit eines geeigneten Betriebssystems und einer Hardwareplattform (nahezu) unumgänglich. Um eine Laufzeitumgebung verfügbar zu halten, gibt es zwei Möglichkeiten. Zum einen kann die Originalhardware aufbewahrt werden (vgl. hierzu Kapitel 12.4 Computermuseum). Zum anderen kann die ursprüngliche Laufzeitumgebung emuliert werden (vgl. hierzu Kapitel 12.3 Emulation). Es existieren bereits unterschiedliche Emulatoren für Hardwareplattformen[2] und Betriebssysteme.

Speicherung und Bereitstellung des digitalen Objekts

Aber nicht nur die Interpretierbarkeit der Informationsobjekte erfordert eine passende Umgebung. Bereits auf der Ebene des Bitstream-Erhalts wird neben dem Speichermedium auch eine Umgebung vorausgesetzt, die das Medium ausliest und die Datenströme an die Darstellungsschicht weitergibt. So brauchen Magnetbänder, CD-ROMs oder DVDs entsprechende Laufwerke und zugehörige Treiber- und Verwaltungssoftware. Bei einer Festplatte sind passende Speicherbusse und ein Betriebssystem, das die Formatierung des eingesetzten Dateisystems verwalten kann, erforderlich.

Literatur

Van Diessen, Raymond J. (2002): preservation requirements in a deposit system. Amsterdam: IBM Netherlands. S. 7-15. http://www-05.ibm.com/nl/dias/resource/preservation.pdf [2007, 20. August]

2 Als Beispiel für die Emulation einer Rechnerarchitektur kann „Dioscuri“ genannt werden. Dioscuri ist eine Java-basierte Emulationssoftware für x86-Systeme. http://dioscuri.sourceforge.net/ [2007, 20.August]

11.2 Digitale Speichermedien

Dagmar Ullrich

Abstract

Datenträger, egal ob analog oder digital, sind nur begrenzt haltbar und müssen früher oder später ausgewechselt werden, um Informationsverlust zu verhindern. Digitale Datenträger veralten in der Regel wesentlich schneller als übliche analoge Medien. Zudem hängt ihre Lesbarkeit von der Verfügbarkeit funktionstüchtiger Lesegeräte ab. Zu den gängigen digitalen Speichermedien zählen Festplatten, Magnetbänder und optische Medien wie CD-ROM oder DVD. Die Unterschiede in Haltbarkeit und Speichereigenschaften entscheiden darüber, in wie weit und in welcher Kombination sie für die Langzeitarchivierung eingesetzt werden können.

Gliederung

Lebensdauer von Trägermedien
Die wichtigsten digitalen Speichermedien
Speichermedien in der Langzeitarchivierung

Lebensdauer von Trägermedien

Um Informationen über die Zeit verfügbar zu halten, müssen sie auf einem zuverlässigen Trägermedium vorliegen. Die Haltbarkeit des Trägermediums ist von wesentlicher Bedeutung für die Verfügbarkeit der Information. Seine begrenzte Lebensdauer erfordert ein rechtzeitiges Übertragen auf ein neues Medium. Mündlich tradierte Gedächtnisinhalte werden durch Auswendiglernen von einer Generation an die nächste weitergereicht. Schriftstücke wie Urkunden, Bücher oder Verträge werden bei Bedarf durch Kopieren vor dem Verfall des Trägermediums geschützt. Auch digitale Daten benötigen Trägermedien, die erhalten und ggf. erneuert werden müssen.[3] Im Vergleich zu herkömmlichen analogen Datenträgern sind digitale Datenträger jedoch in der Regel deutlich kurzlebiger. Neben ihrer Kurzlebigkeit spielt für digitale Datenträger noch ein

3 Der Nachweis der Authentizität ist bei analogem Material wesentlich stärker als bei digitalen Daten an das Trägermedium gebunden. Bei Kopiervorgängen muss dies berücksichtigt werden. Vgl. hierzu Kapitel 8.1.

weiterer Aspekt eine Rolle: Es wird eine Nutzungsumgebung benötigt, um die Datenobjekte zugänglich zu machen. Um ein digitales Trägermedium, z.B. ein Magnetband oder eine CD-ROM lesen zu können, ist ein entsprechendes Laufwerk und die zugehörige Treibersoftware nötig. Wenn man von der Lebensdauer eines digitalen Datenträgers spricht, muss dabei stets auch die Verfügbarkeit der entsprechenden Nutzungsumgebung (Lesegerät und Betriebssystem mit Treibersoftware) im Auge behalten werden. Eine CD-ROM ohne Laufwerk enthält verlorene Daten, selbst wenn die CD-ROM völlig intakt ist.

Die wichtigsten digitalen Speichermedien

In den folgenden Kapiteln werden die drei wichtigsten digitalen Speichermedien, nämlich Festplatte, Magnetbänder und optische Medien vorgestellt. Die genannten Trägermedien lassen sich in zwei Gruppen einteilen: magnetische Medien wie Festplatten und Magnetbänder und optische Medien wie CD-ROM oder DVD. Eine andere mögliche Gruppierung unterscheidet nach Online- und Offline-Speicher. Festplatten werden als Online-Speicher bezeichnet, da sie in der Regel konstant eingeschaltet und für den Zugriff verfügbar sind, Offline-Speichermedien (Magnetbänder, CD-ROM, DVD) dagegen werden nur im Bedarfsfall in ein Laufwerk eingelegt und ausgelesen. Eine dritte mögliche Einteilung der drei Medientypen trennt Medien mit Direktzugriff von so genannten sequentiellen Medien. Beim Direktzugriff kann ein Schreib-/Lesekopf direkt über der gesuchten Stelle positioniert werden. Beim sequentiellen Zugriff muss einer Schreib-/Lesespur gefolgt werden, bis der relevante Abschnitt erreicht wurde. Festplatten arbeiten mit Direktzugriff. Magnetbänder sind dagegen sequentielle Medien. Durch die Online-Verfügbarkeit und den Direktzugriff ist die Festplatte nach wie vor das schnellste der drei gängigen Speichermedien. Dafür ist sie derzeit noch das verschleißanfälligste und teuerste Speichermedium.[4] Die genannten Medientypen werden oft in Kombination eingesetzt. Dabei werden die Medien so angeordnet, dass teure und performante Medien, zumeist Festplatten, Daten mit hoher Zugriffshäufigkeit vorhalten, weniger oft angeforderte Daten dagegen auf preiswerte Offline-Medien ausgelagert werden. Eine solche Anordnung von Speichermedien wird auch als „Hierarchisches Speichermanagement“ (HSM) bezeichnet. Eine entsprechende Empfehlung findet

4 Ob Festplatten immer noch teurer sind als Bandspeicher ist eine derzeit viel diskutierte Frage. Eine interessante Untersuchung findet sich in: McAdam, Dianne (2005): Is Tape Realy Cheaper Than Disk?. White Paper. Nashua: Data Mobility Group. http://www-03.ibm.com/industries/media/doc/content/bin/DMG_tape_disk.pdf?g_type=pspot [2007, 20.August]

sich in Calimera Guidelines for Digital Preservation:
Strategies for both online and offline storage will be needed. Delivery files in continual use will need to be stored online, on servers. Master files are best stored offline since they are less frequently accessed.[5]
Bei größeren Unternehmen und Rechenzentren werden die unterschiedlichen Speichermedien zu umfangreichen Speichernetzwerken zusammengeschlossen. Die verschiedenen Arten von Speichernetzwerken ermöglichen eine gut skalierbare, redundante Speicherung auf unterschiedlichen Medien. In den meisten Fällen kommen hierfür gängige Backup- oder Spiegelungsmechanismen in lokalen Speichernetzwerken zum Einsatz. Andere Konzepte sehen das Zusammenwirken räumlich weit voneinander entfernter Speicherkomponenten vor. Hierzu gehören auch Peer-to-Peer-Netzwerke, wie sie z.B. von der Open Source Software „Lots of Copies Keep Stuff Safe" (LOCKSS)[6] eingesetzt werden.
Speichermedien in der Langzeitarchivierung
Die nachstehende Tabelle vergleicht Festplatte, Bandspeicher und Optische Medien hinsichtlich ihrer Eignung für unterschiedliche Archivierungszeiträume.[7]

Anforderung	**Disk**	**Bandspeicher**	**Optische Medien**
Häufiger Zugriff	Y	N	N
Schnelle Zugriffszeit	Y	N	Vielleicht
Kurze Archivierung (< 1 Jahr)	Y	Y	Y
Mittlere Archivierung (< 10 Jahre)	N	Y	Y
Lange Archivierung (< 20 Jahre)	N	Y	Y
Auslagerung	N	Y	Y
Unveränderbar	N	mit WORM Tape	mit WORM Disc

Die Eignung eines Speichermediums hängt von den Nutzungsanforderungen und ggf. seiner Kombination mit anderen Speichermedien ab. In diesem Sinne gibt es kein für die Langzeitarchivierung in besonderer Weise geeignetes Speichermedium. Vielmehr empfiehlt es sich, eine Speicherstrategie aufzustellen, die

5 o.V. (o.J.) *Digital preservation.* Calimera Guidelines. S.6. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf [2007, 20.August]

6 http://www.lockss.org/lockss/Home [2007, 20.August]

7 Arbeitgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag. S. 45.

den unterschiedlichen Anforderungen der Archivdaten und der durchschnittlichen Lebensdauer der eingesetzten Speichertechniken gerecht werden kann.

Literatur

McAdam, Dianne (2005): Is Tape Really Cheaper Than Disk?. White Paper. Nashua: Data Mobility Group.

http://www-03.ibm.com/industries/media/doc/content/bin/DMG_tape_disk.pdf?g_type=pspot [2007, 20.August]

o.V. (o.J.) Digital preservation. Calimera Guidelines. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf [2007, 20.August]

Arbeitgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit. Eschborn: AWV-Eigenverlag.

11.2.1 Magnetbänder

Dagmar Ullrich

Abstract

Magnetbänder speichern Daten auf einem entsprechend beschichteten Kunststoffband. Dabei können zwei unterschiedliche Verfahren eingesetzt werden, das Linear-Verfahren oder das Schrägspur-Verfahren. Gängige Bandtechnologien verfügen über Funktionen zur Datenkompression und Kontrollverfahren zur Sicherung der Datenintegrität. Die wichtigsten aktuellen Bandtechnologien werden im Überblick vorgestellt. Als Lesegeräte können Einzellaufwerke, automatische Bandwechsler oder umfangreiche Magnetband-Bibliotheken dienen. Verschleiß der Magnetbänder und damit ihre Lebensdauer hängen von der Nutzungsweise und Laufwerksbeschaffenheit ab und fallen daher unterschiedlich aus. Die Haltbarkeit hängt darüber hinaus von der sachgerechten Lagerung ab. Regelmäßige Fehlerkontrollen und -korrekturen sind für einen zuverlässigen Betrieb erforderlich. Magnetbänder eignen sich für die langfristige Speicherung von Datenobjekten, auf die kein schneller oder häufiger Zugriff erfolgt, oder für zusätzliche Sicherungskopien.

Gliederung

Funktionsweise von Magnetbändern
Übersicht der wichtigsten Bandtechnologien
Einzellaufwerke und Bandbibliotheken
Verschleiß und Lebensdauer von Magnetbändern und Laufwerken
Magnetbänder in der Langzeitarchivierung

Funktionsweise von Magnetbändern

Die Datenspeicherung erfolgt durch Magnetisierung eines entsprechend beschichteten Kunststoffbandes. Dabei können zwei unterschiedliche Verfahren eingesetzt werden: das Linear-Verfahren und das Schrägspur-Verfahren. Beim Linear-Verfahren wird auf parallel über die gesamte Bandlänge verlaufende Spuren nacheinander geschrieben. Dabei wird das Band bis zum Ende einer Spur in eine Richtung unter dem Magnetkopf vorbeibewegt. Ist das Ende des Bandes erreicht, ändert sich die Richtung, und die nächste Spur wird bearbeitet. Dieses Verfahren wird auch lineare Serpentinenaufzeichnung genannt. Beim Schrägspur-Verfahren (Helical Scan) dagegen verlaufen die Spuren nicht parallel zum

Band, sondern schräg von einer Kante zur anderen. Der rotierende Magnetkopf steht bei diesem Verfahren schräg zum Band. Die wichtigsten Bandtechnologien, die auf dem Linear-Verfahren beruhen, sind „**L**inear **T**ape **O**pen" (LTO), „**D**igital **L**inear **T**ape (DLT), die Nachfolgetechnologie Super-DLT und „**A**dvanced **D**igital **R**ecording" (ADR). Für das Schrägspurverfahren können als wichtigste Vertreter „Advanced Intelligent Tape" (AIT), Mammoth-Tapes, „**D**igital **A**udio **T**apes" (DAT) und „**D**igital **T**ape **F**ormat" (DTF) genannt werden. Die jeweiligen Technologien nutzen verschiedene Bandbreiten. Gängige Bandformate sind 4 mm, 8 mm, ¼ Zoll (6,2 mm) und ½ Zoll (12,5 mm). Die Kapazitäten liegen im Gigabyte-Bereich mit aktuellen Maximalwerten bei bis zu 1,6 Terabyte (LTO4, mit Datenkompression). Ebenso wie die Bandkapazität hat sich auch die erreichbare Transferrate in den letzten Jahren stark erhöht. Die meisten Bandtechnologien nutzen Datenkompressionsverfahren, um die Kapazität und die Geschwindigkeit zusätzlich zu steigern. Diese Entwicklung wird durch den Konkurrenzdruck immer preiswerteren Festplattenspeichers gefördert. Zur Sicherung der Datenintegrität verfügen die meisten Bandtechnologien über Kontrollverfahren, die sowohl beim Schreiben als auch bei jedem Lesezugriff eingesetzt werden.

Übersicht der wichtigsten Bandtechnologien

Die nachstehende Tabelle listet die oben genannten Technologien im Überblick.[8] Es wurden bewusst auch auslaufende Technologien in die Tabelle aufgenommen (ADR, DTF). Das hat drei Gründe: Erstens werden diese Technologien noch vielerorts eingesetzt, zweitens erlauben die älteren Angaben eine anschauliche Darstellung des Kapazitäts- und Performance-Wachstums in den letzten Jahren und drittens zeigt sich hier, wie schnell Bandtechnologien veraltern und vom Markt verschwinden, auch wenn die Medien selbst eine wesentlich längere Lebensdauer haben.

8 Die Tabelle wurde entnommen und modifiziert aus: Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag. S. 71. Wo erforderlich, sind die Angaben über die Webseiten der Hersteller aktualisiert worden.

Technologie	Aktuelle Version	Kapazität ohne Kompression	Transferrate (MB/sec)	Verfahren	Bandformat	Weiterführende Informationen [18.08.2007]
ADR	ADR 2	60 GB	4	Linear	8 mm	www.speicherguide.de
AIT	AIT-4	200 GB	24	Helical Scan	8 mm	www.aittape.com
DAT	DAT-72	36 GB	6	Helical Scan	4 mm	www.datmgm.com
DLT	DLT-V4	160 GB	10	Linear	½ Zoll	www.dlttape.com
DTF	DTF-2	200 GB	24	Helical Scan	½ Zoll	www.speicherguide.de
LTO-Ultrium	LTO-4	800 GB	160	Linear	½ Zoll	www.lto.org
Mammoth	M2	40 GB	12	Helical Scan	8 mm	www.speicherguide.de
S-DLT	SDLT 600A	300 GB	36	Linear	½ Zoll	www.dlttape.com

Zu ADR siehe 9. Zu DTF siehe 10. Zu Mammoth siehe 11.

Einzellaufwerke und Bandbibliotheken

Magnetbänder werden für Schreib- und Lesevorgänge in ihre zugehörigen Bandlaufwerke eingelegt. Bei kleineren Unternehmen werden in der Regel Einzellaufwerke eingesetzt. Sie werden im Bedarfsfall direkt an einen Rechner angeschlossen und das Einlegen des Bandes erfolgt manuell. Bei steigender Datenmenge und Rechnerzahl kommen automatische Bandwechsler zum Einsatz. Diese Erweiterungen können beliebig skalierbar zu umfangreichen Bandroboter-Systemen (Bandbibliotheken) ausgebaut werden, die über eine Vielzahl von Laufwerken und Bandstellplätzen verfügen. Solche Bandbibliotheken erreichen

9 Die Herstellerfirma OnStream hat 2003 Konkurs anmelden müssen, sodass die Fortführung dieser Technologie unklar ist.

10 Die DTF-Technologie wird seit 2004 nicht fortgeführt.

11 Die Herstellerfirma Exabyte wurde 2006 von Tandberg Data übernommen. Seitdem wird das Mammoth-Format nicht weiterentwickelt.

Ausbaustufen im Petabyte-Bereich.

Verschleiß und Lebensdauer von Magnetbändern und Laufwerken

Die Lebensdauer von Magnetbändern wird üblicherweise mit 2 - 30 Jahre angegeben. Die Autoren von „Speichern, Sichern und Archivieren auf Bandtechnologie" geben sogar eine geschätzte Lebensdauer von mindestens 30 Jahren an:
Für die magnetische Datenspeicherung mit einer 50-jährigen Erfahrung im Einsatz als Massenspeicher kann man sicherlich heute mit Rückblick auf die Vergangenheit unter kontrollierten Bedingungen eine Lebensdauerschätzung von mindestens 30 Jahren gewährleisten.[12]
Die große Spannbreite der Schätzungen erklärt sich durch die unterschiedlichen Bandtechnologien. Auch äußere Faktoren wie Lagerbedingungen und Nutzungszyklen spielen eine wesentliche Rolle für die Haltbarkeit. Da Magnetbänder stets ein passendes Laufwerk benötigen, hängt ihre Lebensdauer auch von der Verfügbarkeit eines funktionstüchtigen Laufwerks ab. Ein schadhaftes Laufwerk kann ein völlig intaktes Band komplett zerstören und somit zu einem Totalverlust der gespeicherten Daten führen. Magnetbänder sollten kühl, trocken und staubfrei gelagert werden. Nach einem Transport oder anderweitiger Zwischenlagerung sollten sie vor Einsatz mind. 24 Stunden akklimatisiert werden. Neben der Lagerung spielt der Einsatzbereich eines Magnetbandes mit der daraus resultierenden Anzahl an Schreib- und Lesevorgängen eine Rolle. Je nach Bandtechnologie und Materialqualität ist der Verschleiß beim Lesen oder Beschreiben eines Tapes unterschiedlich hoch. Auch der Verlauf von Lese- oder Schreibvorgängen beeinflusst die Haltbarkeit der Bänder und Laufwerke. Werden kleine Dateneinheiten im Start-Stopp-Verfahren auf das Magnetband geschrieben, mindert das nicht nur Speicherkapazität und Geschwindigkeit, sondern stellt auch eine wesentlich höhere mechanische Beanspruchung von Bändern und Laufwerken dar. Aus diesem Grund bieten neuere Technologien eine anpassbare Bandgeschwindigkeit (ADR) oder den Einsatz von Zwischenpuffern. Laufwerke die einen ununterbrochenen Datenfluss ermöglichen, werden auch Streamer, die Zugriffsart als Streaming Mode bezeichnet.
Da den Lebensdauerangaben von Herstellern bestimmte Lagerungs- und Nutzungsvoraussetzungen zugrunde liegen, sollte man sich auf diese Angaben nicht ohne weiteres verlassen. Eine regelmäßige Überprüfung der Funktionstüchtigkeit von Bändern und Laufwerken ist in jedem Fall ratsam. Einige Band-

12 Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag. S.85

technologien bringen Funktionen zur Ermittlung von Fehlerraten bei Lesevorgängen und interne Korrekturmechanismen mit. Aus diesen Angaben können Fehlerstatistiken erstellt werden, die ein rechtzeitiges Auswechseln von Medien und Hardware ermöglichen.
Trotz der verhältnismäßig langen Lebensdauer von Magnetbändern und deren Laufwerken sollte nicht übersehen werden, dass die eingesetzten Technologien oft wesentlich kürzere Lebenszyklen haben. Wie bereits oben aus der Tabelle hervorgeht, verschwinden Hersteller vom Markt oder die Weiterentwicklung einer Produktfamile wird aus anderen Gründen eingestellt. Zwar wird üblicherweise die Wartung vorhandener Systeme angeboten, oft aber mit zeitlicher Begrenzung. Aber auch bei der Weiterentwicklung einer Produktfamilie ist die Kompatibilität von einer Generation zur nächsten nicht selbstverständlich. Nicht selten können z.B. Laufwerke einer neuen Generation ältere Bänder zwar lesen, aber nicht mehr beschreiben. Das technische Konzept für die Datenarchivierung des Bundesarchivs sieht daher folgendes vor:
Es sollen nur Datenträger verwendet werden, für die internationale Standards gelten, die am Markt eine ausgesprochen weite Verbreitung haben, als haltbar gelten und daher auch in anderen Nationalarchiven und Forschungseinrichtungen eingesetzt werden. Mit diesen Grundsätzen soll das Risiko minimiert werden, dass der gewählte Archiv-Datenträger vom Markt verschwindet bzw. überraschend von einem Hersteller nicht mehr produziert wird und nicht mehr gelesen werden kann, weil die Laufwerke nicht mehr verfügbar sind.[13]

Magnetbänder in der Langzeitarchivierung

Magnetbänder sind durch ihre vergleichsweise lange Haltbarkeit für die Langzeitarchivierung digitaler Datenbestände gut geeignet. Dies gilt allerdings nur dann, wenn die Daten in dem gespeicherten Format lange unverändert aufbewahrt werden sollen und die Zugriffzahlen eher gering ausfallen. Sind hohe Zugriffszahlen zu erwarten oder ein kurzer Formatmigrationszyklus sollten Bänder in Kombination mit schnellen Medien wie Festplatten zum Speichern von Sicherungskopien eingesetzt werden.

13 Rathje, Ulf (2002): *Technisches Konzept für die Datenarchivierung im Bundesarchiv*. In: Der Archivar, H. 2, Jahrgang 55, S.117-120. (Zitat S. 119).

Literatur

Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit. Eschborn: AWV-Eigenverlag.

Rathje, Ulf (2002): Technisches Konzept für die Datenarchivierung im Bundesarchiv. In: Der Archivar, H. 2, Jahrgang 55, S.117-120.

11.2.2 Festplatten

Dagmar Ullrich

Abstract

Festplatten sind magnetische Speichermedien. Sie speichern Daten mittels eines Schreib-/Lesekopfes, der über drehenden Platten direkt positioniert wird. Die wichtigsten Speicherbusse (S)-ATA, SCSI, SAS und Fibre Channel werden vorgestellt. Festplatten können einzeln oder im Verbund als Speichersubsysteme genutzt werden. Unterschiedliche Speicherkomponenten können komplexe Speichernetzwerke bilden. Die Lebensdauer von Festplatten wird üblicherweise zwischen 3 und 10 Jahren geschätzt. Umgebungseinflüsse wie magnetische Felder, Stöße oder Vibrationen, aber auch Betriebstemperatur und Nutzungszyklen beeinflussen die Haltbarkeit von Festplatten. Festplatten eignen sich für Kurzzeitarchivierung bzw. in Kombination mit anderen Medien zur Verbesserung von Zugriffszeiten. Für eine revisionssichere Archivierung kommen sie in „Content Addressed Storage-Systemen" zum Einsatz, die über Inhalts-Hashes die Datenauthentizität sicherstellen.

Gliederung

Funktionsweise und Speicherbusse
Einzelfestplatten und Festplattensubsysteme
Ausfallursachen und Lebensdauer von Festplatten
Festplatten in der Langzeitarchivierung
Revisionssichere Archivierung mit Content Addressed Storage-Systemen (CAS)

Funktionsweise und Speicherbusse

Festplatten speichern Daten durch ein magnetisches Aufzeichnungsverfahren. Die Daten werden im direkten Zugriff (random access) von einem positionierbaren Schreib-/Lesekopf auf die rotierenden Plattenoberflächen geschrieben bzw. von dort gelesen. Festplatten können beliebig oft beschrieben und gelesen werden. Die aktuelle Maximalkapazität einer einzelnen Festplatte liegt bei einem Terabyte. Festplatten zeichnen sich gegenüber sequentiellen Medien wie Magnetbändern durch schnellen Zugriff auf die benötigten Informationsblöcke aus. Die Zugriffsgeschwindigkeit einer Festplatte hängt vor allem von der Positionierzeit des Schreib-/Lesekopfes, der Umdrehungsgeschwindigkeit

der Platten und der Übertragungsrate, mit der die Daten von/zur Platte übertragen werden, ab. Die Übertragungsrate wird wesentlich von der Wahl des Speicherbusses, der Anbindung der Festplatte an den Systembus, bestimmt. Die Speicherbusse lassen sich in parallele und serielle Busse unterscheiden. Die Entwicklung paralleler Busse ist rückläufig, da bei zunehmender Übertragungsrate die Synchronisation der Datenflüsse immer schwieriger wird. Die wichtigsten Standards für Speicherbusse sind: „Advanced Technology-Attachment" (ATA). Dieser ursprünglich parallele Bus wird heute fast ausschließlich seriell als S-ATA eingesetzt. „Small Computer Systems Interface" (SCSI) wurde ebenfalls ursprünglich als paralleler Bus entwickelt und wird heute vorwiegend seriell als Serial-Attached-SCSI (SAS) betrieben. Dieses Bussystem zeichnet sich durch hohe Übertragungsraten und einfache Konfiguration aus. Fibre Channel[14] (FC) ist ein originär serieller Bus. Er ermöglicht die Hochgeschwindigkeitsübertragung großer Datenmengen und die Verbindung von Speicherkomponenten mit unterschiedlichen Schnittstellen. Er kommt daher hauptsächlich bei größeren Speichersubsystemen oder komplexen Speichernetzwerken zum Einsatz.
Festplatten werden häufig nach ihren Schnittstellen als (S-)ATA-, SCSI- oder SAS-Platten bezeichnet. SCSI- oder SAS-Platten bieten schnelle Zugriffszeiten, sind jedoch im Vergleich zu S-ATA-Platten teuer. S-ATA-Platten dienen vorwiegend dem Speichern großer Datenmengen mit weniger hohen Zugriffsanforderungen. Die ursprünglich aus dem Notebook-Umfeld stammende, heute zunehmend aber auch als mobiles Speichermedium z.B. für Backup-Zwecke eingesetzte USB-Platte basiert derzeit intern meist auf einer Platte mit (S)-ATA-Schnittstelle.

Einzelfestplatten und Festplattensubsysteme

Festplatten können intern in PCs oder Servern eingebaut oder auch als extern angeschlossener Datenspeicher eingesetzt werden. Die Kapazität einzelner Platten kann durch ihren Zusammenschluss zu Speichersubsystemen (Disk-Arrays) bis in den Petabyte-Bereich[15] erweitert werden. Solche Speichersubsysteme werden meist als RAID-Systeme bezeichnet. RAID steht für „Redundant Array of Independent[16] Disks. "Redundant" weist hier auf den wichtigsten Einssatz-

14 Die Bezeichnung Fibre Channel kann insofern irreführend sein, als dass dieser serielle Speicherbus sowohl mit Glasfaser als auch mittels herkömmlicher Kupferkabel umgesetzt werden kann.

15 Die Bezeichnung Fibre Channel kann insofern irreführend sein, als dass dieser serielle Speicherbus sowohl mit Glasfaser als auch mittels herkömmlicher Kupferkabel umgesetzt werden kann.

16 Da RAID-Systeme die Möglichkeit bieten, auch preiswerte Festplatten mit hoher Ausfallsi-

zweck dieser Systeme hin: Der Zusammenschluss von Einzelplatten dient nicht nur der Kapazitätserweiterung, sondern vorwiegend der verbesserten Ausfallsicherheit und Verfügbarkeit. Die Platten in RAID-Systemen können so konfiguriert werden, dass bei Ausfall einzelner Platten die betroffenen Daten über die verbliebenen Platten im laufenden Betrieb rekonstruiert werden können. In RAID-Systemen kommen üblicherweise SCSI-Platten zum Einsatz. Zunehmend werden aus Kostengründen auch (S-)ATA-Platten eingesetzt, wobei das Subsystem selbst über SCSI oder FC mit dem Speichernetzwerk verbunden wird. Interessant mit Blick auf ihre Langlebigkeit sind die verhältnismäßig neuen MAID-Systeme. MAID steht für „Massive Array of Idle Disks". Im Unterschied zu herkömmlichen Festplatten-RAIDs sind die Platten dieser Speicher-Arrays nicht konstant drehend, sondern werden nur im Bedarfsfall aktiviert. Dies mindert den Verschleiß ebenso wie Stromverbrauch und Wärmeentwicklung, kann aber zu Einbußen in der Zugriffsgeschwindigkeit führen.

Ausfallursachen und Lebensdauer von Festplatten

Die Lebensdauer von Festplatten wird sehr unterschiedlich eingeschätzt. Zumeist wird eine Lebensdauer zwischen 3 und 10 Jahren angenommen. Es finden sich jedoch auch wesentlich höhere Angaben von bis zu 30 Jahren. In der Regel werden als Haupteinflüsse die Betriebstemperatur und der mechanische Verschleiß angesehen. Die übliche Betriebstemperatur sollte bei 30°-45°C liegen, zu hohe, aber auch sehr niedrige Temperaturen können der Festplatte schaden. Ein mechanischer Verschleiß ist bei allen beweglichen Teilen möglich. So sind die Lager der drehenden Platten und der bewegliche Schreib-/Lesekopf bei hohen Zugriffszahlen verschleißgefährdet. Die Gefahr, dass Platten durch lange Ruhezeiten beschädigt werden („sticky disk"), ist bei modernen Platten deutlich verringert worden. Zwei Risiken sind bei Festplatten besonders ernst zu nehmen, da sie einen Totalverlust der Daten bedeuten können: zum einen der so genannte Head-Crash. Ein Head-Crash bedeutet, dass der Schreib-/Lesekopf die drehenden Platten berührt und dabei die Plattenbeschichtung zerstört. Zum anderen können umgebende Magnetfelder die magnetischen Aufzeichnungen schädigen. Festplatten sollten daher in einer Umgebung aufbewahrt werden, die keine magnetischen Felder aufweist, gleichmäßig temperiert ist und die Platte keinen unnötigen Stößen oder sonstigen physischen Beeinträchtigungen aussetzt. In welchem Maße die unterschiedlichen Einflüsse die Lebensdauer von Festplatten beeinträchtigen, wird üblicherweise durch Extrapolation von Labortests festgelegt. Hieraus resultieren die Herstellerangaben zu Lebensdauer

cherheit zu betreiben, wird das „I" in RAID auch mit „inexpensive" übersetzt.

und Garantiezeiten. Die Lebensdauer einer Festplatte wird üblicherweise mit „mean time before failure“ (MTBF) angegeben. Diese Angabe legt die Stunden fest, die eine Platte betrieben werden kann, bevor Fehler zu erwarten sind. Die Betriebsdauer sollte sich jedoch nicht nur an der MTBF ausrichten, da im Produktivbetrieb oft deutliche Abweichungen von diesen Werten feststellbar sind. Es empfiehlt sich stets auch der Einsatz und die Weiterentwicklung von Überwachungssoftware.

Festplatten in der Langzeitarchivierung

Welche Rolle kann ein Medium, dem eine durchschnittliche Lebensdauer von 5 Jahren zugesprochen wird, für die Langzeitarchivierung von digitalen Datenbeständen spielen? Als Trägermedium zur langfristigen Speicherung von Daten sind langlebigere Medien wie Magnetbänder nicht nur aufgrund ihrer Lebensdauer, sondern auch aus Kostengründen in der Regel besser geeignet. Festplatten können aber in zwei möglichen Szenarien auch für Langzeitarchivierungszwecke sinnvoll sein. Zum einen können sie die Zugriffszeiten auf Archivinhalte deutlich verbessern, wenn sie in Kombination mit anderen Medien in einem hierarchischen Speichermanagement eingesetzt werden. Zum anderen können beispielsweise Formatmigrationen schon nach kurzer Zeit für einen Teil der Archivobjekte erforderlich werden. In diesem Fall ist eine langfristige Speicherung der Dateien gar nicht erforderlich, sondern viel eher deren zeitnahes Auslesen und Wiedereinstellen nach erfolgter Formataktualisierung. Die veralteten Originalversionen können dann auf ein langlebiges Medium ausgelagert werden. Für die jeweils aktuellen Versionen jedoch, die möglicherweise einen kurzen Formatmigrationszyklus haben, kann eine Festplatte ein durchaus geeignetes Trägermedium sein.

Revisionssichere Archivierung mit Content Addressed Storage-Systemen (CAS)

In Wirtschaftsunternehmen und im Gesundheitswesen sind die Anforderungen an Archivierungsverfahren oft an die Erfüllung gesetzlicher Auflagen gebunden. Zu diesen Auflagen gehört oft der Nachweis der Datenauthentizität. Eine Möglichkeit, diese geforderte Revisionssicherheit herzustellen, liegt in der Verwendung von Speichermedien, die nicht überschrieben werden können. Hierfür wurde in der Vergangenheit auf WORM-Medien (Write Once Read Many) zurückgegriffen. Heute werden CD-ROM oder DVD bevorzugt. Eine Alternative hierzu stellen so genannte CAS-Systeme auf Festplattenbasis dar. CAS-Systeme nutzen gut skalierbare Festplattenspeicher in Kombination mit internen Servern und einer eigenen Verwaltungssoftware. Das Grundprinzip beruht auf der

Erstellung von Checksummen bzw. Hashes zu jedem eingestellten Inhalt. Über diese Inhalts-Hashes werden die Objekte adressiert. Der Hash-Wert sichert dabei die Authentizität des über ihn adressierten Inhalts. Dieses Verfahren ist an die Verfügbarkeit des CAS-Systems und der Funktionstüchtigkeit der eingesetzten Hardware gebunden. In der Regel können einzelne Komponenten im laufenden Betrieb ausgetauscht und aktualisiert werden.

12 Digitale Erhaltungsstrategien

Einleitung

Stefan E. Funk

Wie lassen sich die Dinge bewahren, die uns wichtig sind, Objekte, die wir der Nachwelt am allerliebsten in genau dem Zustand, in dem sie uns vorliegen, erhalten wollen?

Handelt es sich bei diesen Objekten um Texte oder Schriften, wissen wir, dass Stein- und Tontafeln sowie Papyri bei geeigneter Behandlung mehrere tausend Jahre überdauern können. Auch bei Büchern haben wir in den letzten Jahrhunderten Kenntnisse darüber gesammelt, wie diese zu behandeln sind bzw. wie diese beschaffen sein müssen, um nicht der unfreiwilligen Zerstörung durch zum Beispiel Säurefraß oder Rost durch eisenhaltige Tinte anheim zu fallen. Auch Mikrofilme aus Cellulose mit Silberfilm-Beschichtung sind bei richtiger Lagerung viele Jahrzehnte, vielleicht sogar Jahrhunderte, haltbar. Alle diese Medien haben den Vorteil, dass sie, wenn sie als die Objekte, die sie sind, erhalten werden können, von der Nachwelt ohne viele Hilfsmittel interpretiert werden können. Texte können direkt von Tafeln oder aus Büchern gelesen und Mikrofilme mit Hilfe eines Vergrößerungsgerätes recht einfach lesbar gemacht werden.

Bei den digitalen Objekten gibt es zwei grundlegende Unterschiede zu den oben genannten analogen Medien: Zum einen werden die digitalen Informationen als Bits (auf Datenträgern) gespeichert. Ein Bit ist eine Informationseinheit und hat entweder den Wert „0“ oder den Wert „1“. Eine Menge dieser Nullen und Einsen wird als Bitstream bezeichnet. Die Lebensdauer der Bits auf diesen Datenträgern kennen wir entweder nur aus Laborversuchen oder wir haben noch nicht genug Erfahrungswerte für eine sichere Angabe der Lebensdauer über einen langen Zeitraum hinweg sammeln können. Schließlich existieren diese Datenträger erst seit einigen Jahren (bei DVDs) oder Jahrzehnten (bei CDs). Eine Reaktion auf die Unsicherheit über die Lebensdauer dieser Medien ist die Bitstreamerhaltung sowie die Mikroverfilmung.

Zum anderen ist keines der digitalen Objekte ohne technische Hilfsmittel nutzbar. Selbst wenn wir die Nullen und Einsen ohne Hilfsmittel von den Medien lesen könnten, dann könnten wir wenig bis gar nichts mit diesen Informationen anfangen. Da diese konzeptuellen Objekte digital kodiert auf den Medien gespeichert sind, bedarf es oben genannter Hilfsmittel, die diese Informationen interpretieren können. Als Hilfsmittel dieser Art ist einerseits die Hardware zu sehen, die die Daten von den Medien lesen kann (beispielsweise CD- bzw. DVD-Laufwerke) und natürlich die Computer, die diese Daten weiterverarbeiten. Andererseits wird die passende Software benötigt, die die Daten interpretiert und so die digitalen Objekte als konzeptuelle Objekte erst oder wieder nutzbar macht.

Kann der Bitstream nicht mehr interpretiert werden, weil das Wissen um eine korrekte Interpretation verloren ging, ist der Inhalt des konzeptuellen Objektes verloren, obwohl die eigentlichen Daten (der Bitstream) noch vorhanden sind. Lösungsansätze für dieses Problem sind die Migration und die Emulation. Eine weitere Idee ist es, in einem so genannten Computermuseum die originale Hard- und Software bereitzustellen und so die konzeptuellen Objekte zu erhalten.

12.1 Bitstream Preservation

Dagmar Ullrich

Abstract

Grundlage aller Archivierungsaktivitäten ist der physische Erhalt der Datenobjekte, die Bitstream[1] Preservation. Es wird eine Speicherstrategie vorgeschlagen, die auf einer redundanten Datenhaltung auf mindestens zwei unterschiedlichen, marktüblichen und standardisierten Speichertechniken basiert. Die eingesetzten Speichermedien sollten regelmäßig durch aktuelle ersetzt werden, um sowohl dem physischen Verfall der Speichermedien als auch dem Veralten der eingesetzten Techniken vorzubeugen. Es werden vier Arten von Migrationsprozessen vorgestellt. Das sind: Refreshment, Replication, Repackaging und Transformation. Als Medienmigration im engeren Sinne werden nur die beiden ersten, Refreshment und Replication, angesehen. Sie bezeichnen das Auswechseln einzelner Datenträger (refreshing) oder eine Änderung eingesetzter Speicherverfahren (replication). Durch die kurzen Lebenszyklen digitaler Speichermedien erfolgt ein Erneuern der Trägermedien oft im Rahmen der Aktualisierung der eingesetzten Speichertechnik.

Gliederung

- Physischer Erhalt der Datenobjekte
- Verfahrensvorschläge für eine Bitstream Preservation
- Redundanz, Speichertechniken und Standards
- Regelmäßige Medienmigration
- Refreshment und Replication
- Zusammenfassung

Physischer Erhalt der Datenobjekte

Um digitale Daten langfristig verfügbar zu halten, muss an zwei Stellen angesetzt werden. Zum einen muss der physische Erhalt des gespeicherten Datenobjekts (Bitstreams) auf einem entsprechenden Speichermedium gesichert

1 Der Begriff „Bitstream“ wird hier als selbsterklärend angesehen. Eine Erläuterung des Begriffs findet sich in: Rothenberg, Jeff (1999): *Ensuring the Longevity of Digital Information*. http://www.clir.org/pubs/archives/ensuring.pdf [2007, 28.August]
Bei diesem Text handelt es sich um eine ausführlichere Fassung eines gleichnamigen Artikels, der 1995 in der Zeitschrift „Scientific American“, Band 272, Nummer 1, Seiten 42-47 erschienen ist.

werden. Zum anderen muss dafür Sorge getragen werden, dass dieser Bitstream auch interpretierbar bleibt, d.h. dass eine entsprechende Hard- und Software-Umgebung verfügbar ist, in der die Daten für einen menschlichen Betrachter lesbar gemacht werden können. Ohne den unbeschädigten Bitstream sind diese weiterführenden Archivierungsaktivitäten sinnlos. Der physische Erhalt der Datenobjekte wird auch als „Bitstream Preservation" bezeichnet. Für den physischen Erhalt des Bitstreams ist eine zuverlässige Speicherstrategie erforderlich.

Verfahrensvorschläge für eine Bitstream Preservation

Die nachstehenden vier Verfahrensvorschläge können als Grundlage für eine zuverlässige Speicherstrategie zur Sicherstellung des physischen Erhalts der archivierten Datenobjekte verwendet werden:[2]

1. *Redundante Datenhaltung*: Die Daten sollten in mehrfacher Kopie vorliegen. Zur Sicherung gegen äußere Einflüsse empfiehlt sich auch eine räumlich getrennte Aufbewahrung der unterschiedlichen Kopien.
2. *Diversität eingesetzter Speichertechnik*: Die Daten sollten auf mindestens zwei unterschiedlichen Datenträgertypen gesichert werden.
3. *Standards*: Die verwendeten Speichermedien sollten internationalen Standards entsprechen und auf dem Markt eine weite Verbreitung aufweisen.
4. *Regelmäßige Medienmigration*: Die verwendeten Speichertechniken bzw. Datenträger müssen regelmäßig durch neue ersetzt werden.

Redundanz, Speichertechniken und Standards

Eine mehrfach *redundante Datenhaltung* ist in vielen Bereichen der Datensicherung üblich. Bei wertvollen, insbesondere bei nicht reproduzierbaren Daten wird man sich nicht auf eine einzige Kopie verlassen wollen. Um das Risiko äußerer Einflüsse wie Wasser- oder Brandschäden zu verringern, bietet sich die räumlich getrennte Aufbewahrung der Kopien an. Um auch die Gefahr eines Datenverlusts durch menschliches Versagen oder Vorsatz einzuschränken, kann eine Aufbewahrung bei zwei unabhängigen organisatorischen Einheiten in das

2 Die Auflistung erhebt keinen Anspruch auf Vollständigkeit. Ähnliche Aufstellungen finden sich z.B. in: Rathje, Ulf (2002): Technisches Konzept für die Datenarchivierung im Bundesarchiv. In: Der Archivar, H. 2, Jahrgang 55, S.117-120. http://www.archive.nrw.de/archivar/2002-02/heft2_02_s117_126.pdf [2007, 28.August] und: o.V. (o.J.) Digital preservation. Calimera Guidelines. S.3. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf [2007, 28.August]

Redundanzszenario mit einbezogen werden. Zusätzliche Sicherheit lässt sich gewinnen, indem die jeweiligen Kopien *auf unterschiedlichen Speichertechniken* gehalten werden. Dies mindert das Risiko eines Datenverlusts durch Veralterung einer der eingesetzten Techniken. Sofern vorhanden, sollten Fehlererkennungs- und Korrekturmechanismen zur Sicherung der Datenintegrität eingesetzt werden. Weiter sollte die Funktionstüchtigkeit der Speichermedien und Lesegeräte anhand von Fehlerstatistiken überwacht werden. Die sachgerechte Handhabung von Datenträgern und Lesegeräten ist in jedem Fall vorauszusetzen. Alle verwendeten Speichertechniken bzw. -medien sollten auf internationalen *Standards* basieren und über eine möglichst breite Nutzergruppe verfügen.

Regelmäßige Medienmigration

Als Medienmigration kann jeder Vorgang betrachtet werden, bei dem das physische Trägermedium eines Datenobjekts innerhalb eines Archivs geändert und der Vorgang mit der Absicht durchgeführt wird, das Datenobjekt zu erhalten, indem die alte Instanz durch die neue ersetzt wird. Eine entsprechende Definition von „Digital Migration" findet sich im OAIS-Referenzmodell[3]:

> *Digital Migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes:*
>
> *- a focus on the Preservation of the full information content*
>
> *- a perspective that the new archival implementation of the information is a replacement for the old; and*
>
> *- full control and responsibility over all aspects of the transfer resides with the OAIS.*

Im OAIS-Referenzmodell werden vier Arten der Migration genannt: Refreshment, Replication, Repackaging und Transformation[4].
Refreshment: Als Refreshment werden Migrationsprozesse bezeichnet, bei denen einzelne Datenträger gegen neue, gleichartige Datenträger ausgetauscht werden. Die Daten auf einem Datenträger werden direkt auf einen neuen Da-

3 Consultative Committee for Space Data Systems (CCSDS) (2002): *Reference Model for an Open Archival Information System (OAIS).* Blue Book. Washington DC. Seite 5-1. vgl. auch Kapitel 7. http://public.ccsds.org/publications/archive/650x0b1.pdf [2007, 19. Februar]

4 Consultative Committee for Space Data Systems (CCSDS) (2002): *Reference Model for an Open Archival Information System (OAIS).* A.a.O. Seite 5-4.

tenträger gleichen Typs kopiert, der anschließend den Platz des alten in der Speicherinfrastrukur des Archivs einnimmt. Weder an den Daten noch an der Speicherinfrastruktur werden also Änderungen vorgenommen, es wird lediglich ein Datenträger gegen einen gleichartigen anderen ausgetauscht.
Replication: Eine Replication ist ein Migrationsprozess, bei dem ebenfalls Daten von einem Datenträger auf einen neuen kopiert werden. Bei der Replication jedoch kann es sich bei dem neuen Datenträger auch um einen andersartigen, z.B. aktuelleren, handeln. Andersartige Datenträger erfordern eine entsprechende Anpassung der Speicherinfrastruktur. Der neue Datenträger kann in der Regel nicht unmittelbar den Platz des alten einnehmen. Der wesentliche Unterschied zum Refreshment liegt daher in den mit dem Kopierprozess einhergehenden Änderungen der verwendeten Speicherinfrastruktur.
Repackaging: Ein Repackaging ist ein Migrationsprozess, bei dem ein sogenanntes. Archivpaket verändert wird. Diese Änderung betrifft nicht die eigentlichen Inhaltsdaten, sondern die Struktur des Archivpakets.
Transformation: eine Transformation ist ein Migrationsprozess, bei dem auch die Inhaltsdaten des Archivpakets verändert werden.
Refreshment und Replication können als Medienmigrationen im engeren Sinne angesehen werden. Der Umkopierprozess erfolgt in beiden Fällen mit der Absicht, das Trägermedium zu ersetzen, unabhängig davon, welche Inhalte auf ihm abgelegt sind. Die Replication wird im Folgenden im Sinne eines Technologiewechsels interpretiert.[5] in Refreshment beschränkt sich dagegen auf den Wechsel einzelner Datenträger innerhalb einer Speichertechnik, z.B. einer Magnetbandgeneration. Bei Repackaging und Transformation dagegen werden auch die Datenobjekte selbst umgeschrieben. Ein Beispiel für ein Repackaging ist die Änderung des Packformats von ZIP zu TAR. Eine Formatmigration, z.B. von JPG zu TIFF, ist dagegen eine Transformation, da die Inhalte des Archivpakets verändert werden. Die Unterscheidung dieser vier Arten von Migrationen erleichtert die begriffliche Abgrenzung einer Medienmigration von einer Formatmigration. Eine Formatmigration umfasst letztlich immer auch eine Medienmigration, da ein neues Datenobjekt erstellt und auf einem eigenen Trägermedium abgelegt wird. Die Formatmigration erfolgt aber mit Blick auf die künftige Interpretierbarkeit des Bitsreams, die Medienmigration im engeren Sinne hingegen dient dessen Erhalt. Für die Bitstream Preservation sind nur die beiden ersten, Refreshment und Replication, wesentlich, da die beiden anderen den Bitstream verändern. Ein Refreshment ist in der Regel weniger aufwendig

5 Eine Replication muss nach der zitierten Definition nicht notwendig von einem veralteten Medium auf ein aktuelleres erfolgen, sondern ggf. auch auf ein gleichartiges. In der Praxis wird das aber eher selten der Fall sein.

als eine Replication, da nicht das Speicherverfahren, sondern nur einzelne Datenträger erneuert werden.

Refreshment und Replication

Ein Erneuern (refreshing) einzelner Datenträger kann aufgrund von Fehlerraten oder auf der Basis bestimmter Kriterien wie Zugriffshäufigkeit oder Alter erfolgen. Der Aufwand solcher Maßnahmen ist gegen die Wahrscheinlichkeit eines Datenverlusts durch einen fehlerhaften Datenträger abzuwägen. Auf der einen Seite können zusätzliche Kontrollverfahren eine sehr hohe Systemlast erzeugen, die den aktiven Zugriff auf die Daten beträchtlich einschränken kann. Zudem sind die Beurteilungskriterien wie Zugriffshäufigkeit, Alter und ggf. die tolerierbare Fehlerrate oft strittig und zum Teil nur mit teurer Spezialsoftware oder auch gar nicht feststellbar. Nicht selten können sie auch im Einzelfall durch Unterschiede in Produktionsablauf oder Handhabung zwischen Datenträgern desselben Typs stark variieren. Auf der anderen Seite wird die Haltbarkeit von Trägermedien aufgrund des raschen Technologiewandels meist gar nicht ausgereizt. Die Wahrscheinlichkeit schadhafter Datenträger durch altersbedingten Verfall ist daher eher gering. Um diesen Zusammenhang deutlich zu machen, kann die durchschnittliche Lebensdauer eines Datenträgers von seiner durchschnittlichen Verfallszeit unterschieden werden.[6]

> *„Medium Expected Lifetime (MEL): The estimated amount of time the media will be supported and will be operational within the electronic deposit system."*
>
> *"Medium Decay Time (MDT): The estimated amount of time the medium should operate without substantial read and write errors."*

Die Definition der durchschnittlichen Lebensdauer enthält zwei durch „und" verbundene Zeitangaben. Die eine bezieht sich auf die Dauer der Unterstützung eines Speichermediums durch den Hersteller, die andere auf die Dauer des Einsatzes eines Speichermediums im digitalen Archiv. Diese beiden Zeitspannen können durchaus differieren. Nicht selten zwingt die wegfallende Unterstützung durch den Hersteller zur Migration, auch wenn die vorhandenen Systeme voll funktionsfähig sind und noch weiter betrieben werden könnten. Für Speichertechniken, die vom Hersteller nicht mehr unterstützt werden, können Ersatzteile oder technische Betreuung nicht mehr garantiert werden. Ihr Weiterbetrieb ist daher nicht ratsam. Der Begriff der durchschnittlichen Le-

6 Van Diessen, Raymond J. und van Rijnsoever, Ben J. (2002): *Managing Media Migration in a Deposit System. IBM/KB Long-Term Preservation Study Report Series Nr. 5.* Amsterdam: IBM Niederlande. S.4. http://www-5.ibm.com/nl/dias/resource/migration.pdf [2007, 28. August]

bensdauer wird aus diesen Gründen hier als die durchschnittlich zu erwartende Hersteller-Unterstützung interpretiert. Solange diese durchschnittliche Lebensdauer unter der durchschnittlichen Verfallszeit liegt, ist ein Ausfall einzelner Datenträgern selten zu erwarten. Statt aufwendiger Kontrollen der Datenträger kann es in diesem Fall einfacher sein, auf eine redundante Datenhaltung zu vertrauen, im konkreten Fehlerfall einzelne Datenträger oder Laufwerke zu ersetzen und den gesamten Bestand im Rahmen eines Technologiewechsels (Replication) komplett auszutauschen.
Eine Replication im Sinne eines Technologiewechsels umfasst Änderungen in der bestehenden Speicherinfrastrukur. Erforderliche Technologiewechsel können sehr unterschiedlich ausfallen. Sie können von einer Magnetbandgeneration zur nächsten reichen oder einen vollständigen Wechsel z.B. von Magnetbändern zu optischen Medien bedeuten. Im ersten Schritt muss die neue Speichertechnik in die bestehende Infrastruktur integriert werden. Anschließend müssen die betroffenen Datenbestände von der alten Technik auf die neue umkopiert werden. Bei großen Datenmengen mit ggf. hohen Sicherheits- oder Verfügbarkeitsansprüchen können diese Umkopierprozesse aufwändig und langwierig sein. Die Lesegeschwindigkeit der älteren Speichermedien wird in der Regel langsamer sein als die Schreibgeschwindigkeit der neuen. Beide müssen für einen Kopierprozess koordiniert werden, ggf. über Zwischenspeicher. Der Übertragungsvorgang muss abgeschlossen sein, bevor die alte Speichertechnik unbrauchbar wird. An diesem Punkt sei auf die oben ausgeführte Interpretation von „Medium Expected Lifetime“ hingewiesen. Dass der Migrationsprozess abgeschlossen sein muss, bevor eine Speichertechnik nicht mehr auf dem Markt ist, wäre ein sehr hoher Anspruch, da viele Speichermedien nur drei bis 5 Jahren lang angeboten werden. Unter Umständen kann ein solcher Anspruch je nach Wert der betroffenen Daten gerechtfertigt sein. Häufig bieten Hersteller die Unterstützung von Speichermedien einige Jahre länger an, als diese Technik aktiv vertrieben wird. Dies verlängert die zuverlässige Einsatzdauer von Speichertechniken. Eine zusätzliche Sicherheit kann in diesem Kontext auch der Verfahrensvorschlag, unterschiedliche Speichertechniken einzusetzen, bieten.

Zusammenfassung

Ein Langzeitarchiv muss über zuverlässige Speicherstrategien verfügen, die nicht nur ein „Refreshment“ eingesetzter Datenträger innerhalb einer Speichertechnik ermöglichen, sondern darüber hinaus auch die Erneuerung ganzer Speichertechniken. Solche Strategien müssen sicherstellen, dass zu keinem Zeitpunkt Datenbestände unzugänglich werden, weil ihre Trägermedien nicht mehr lesbar sind.

Literatur

Rothenberg, Jeff (1999), *Ensuring the Longevity of Digital Information*. http://www.clir.org/pubs/archives/ensuring.pdf [2007, 28.August]. Bei diesem Text handelt es sich um eine ausführlichere Fassung eines gleichnamigen Artikels, der 1995 in der Zeitschrift „Scientific American", Band 272, Nummer 1, Seiten 42-47 erschienen ist.

Rathje, Ulf (2002): *Technisches Konzept für die Datenarchivierung im Bundesarchiv*. In: Der Archivar, H. 2, Jahrgang 55, S.117-120. http://www.archive.nrw.de/archivar/2002-02/heft2_02_s117_126.pdf [2007, 28.August]

o.V. (o.J.) *Digital preservation*. Calimera Guidelines. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf [2007, 28.August]

Consultative Committee for Space Data Systems (CCSDS) (2002): *Reference Model for an Open Archival Information System (OAIS)*. Blue Book. Washington DC. Seite 5-1. http://public.ccsds.org/publications/archive/650x0b1.pdf [2007, 19. Februar]

Van Diessen, Raymond J. und van Rijnsoever, Ben J. (2002): *Managing Media Migration in a Deposit System*. IBM/KB Long-Term Preservation Study Report Series Nr. 5. Amsterdam: IBM Niederlande. http://www-5.ibm.com/nl/dias/resource/migration.pdf [2007, 28. August]

12.2 Migration

Stefan E. Funk

Migration und Emulation

Wenn die Archivierung des Bitstreams sichergestellt ist (siehe Bitstreamerhaltung), kann man beginnen, sich über die Archivierung und vor allem über die Nutzung von digitalen Objekten Gedanken zu machen. Bei nicht digitalen Medien wie Büchern und Mikrofilmen hat man in den letzten Jahrzehnten und Jahrhunderten sehr viel Erfahrung mit deren Erhaltung gesammelt, das heißt, auf physikalischer Ebene konnten und können diese Medien sehr lange verfügbar gehalten werden. Ein Buch braucht als zu erhaltendes Objekt auch nur auf der physischen Ebene betrachtet zu werden, denn zum Benutzen eines Buches reicht es aus, das Buch selbst zu erhalten und so die Lesbarkeit zu gewährleisten.

Zwei Strategien, die die Lesbarkeit der archivierten digitalen Dokumente über lange Zeit (Long Term) garantieren sollen, sind zum einen die Migration und zum anderen die Emulation. „Long term" wird vom Consultative Committee for Space Data Systems (CCSDS) definiert als: „Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely."

Die Migration passt die digitalen Objekte selbst einem neuen Umfeld an, die Dokumente werden zum Beispiel von einem veralteten Dateiformat in ein aktuelles konvertiert. Mit der Emulation wird das originäre Umfeld der digitalen Objekte simuliert, das neue Umfeld also an die digitalen Objekte angepasst. Diese Strategien können alternativ genutzt werden, sie sind unabhängig voneinander.

Um ein digitales Dokument archivieren und später wieder darauf zugreifen zu können, sind möglichst umfassende Metadaten nötig, also Daten, die das digitale Objekt möglichst genau beschreiben. Dazu gehören in erster Linie die technischen Metadaten. Für die Migration sind weiterhin die Provenance Metadaten wichtig, die wie oben erläutert die Herkunft des Objekts beschreiben. Deskriptive Metadaten sind aus technischer Sicht nicht so interessant. Sie werden benötigt, um später einen schnellen und komfortablen Zugriff auf die Objekte zu ermöglichen und rechtliche Metadaten schließlich können genutzt werden, um Einschränkungen für die Migration, die Emulation und den Zugriff auf die

digitalen Objekte festzulegen.

Migration

Mit dem Stichwort Migration werden innerhalb der Langzeitarchivierungs-Community unterschiedliche Prozesse bezeichnet, dies sind sowohl die Datenträgermigration als auch die Daten- oder Formatmigration.

Bei der Datenträgermigration werden Daten von einem Träger auf einen anderen kopiert, z.B. von Festplatte auf CD, von DVD auf Band etc. Diese Art der Migration ist die Grundlage der physischen Erhaltung der Daten, der Bitstream Preservation.

Bei einer Datenmigration (auch Formatmigration genannt) werden Daten von einem Datenformat in ein aktuelleres, möglichst standardisiertes und offen gelegtes Format überführt. Dies sollte geschehen, wenn die Gefahr besteht, dass archivierte Objekte aufgrund ihres Formates nicht mehr benutzt werden können. Das Objekt selbst wird so verändert, dass seine Inhalte und Konzepte erhalten bleiben, es jedoch auf aktuellen Rechnern angezeigt und benutzt werden kann. Problematisch ist bei einer Datenmigration der möglicherweise damit einhergehende Verlust an Informationen. So ist es zum Beispiel möglich, dass sich das äußere Erscheinungsbild der Daten ändert oder - noch gravierender - Teile der Daten verloren gehen.

Eine verlustfreie Migration ist dann möglich, wenn sowohl das Original-Format wie auch das Ziel-Format eindeutig spezifiziert sind, diese Spezifikationen bekannt sind UND eine Übersetzung von dem einen in das andere Format ohne Probleme möglich ist. Hier gilt: Je einfacher und übersichtlicher die Formate, desto größer ist die Wahrscheinlichkeit einer verlustfreien Migration. Bei Migration komplexer Datei-Formate ist ein Verlust an Informationen wahrscheinlicher, da der Umfang einer komplexen Migration nicht unbedingt absehbar ist. Eine Migration eines Commodore-64 Computerspiels in ein heute spielbares Format für einen PC ist sicherlich möglich, jedoch ist es (a) sehr aufwändig, (b) schlecht bzw. gar nicht automatisierbar und (c) das Ergebnis (sehr wahrscheinlich) weit vom Original entfernt.

Beispiel: Alte und neue PCs

- Sie haben einen recht alten PC, auf dem Sie seit langem Ihre Texte schrei-

ben, zum Beispiel mit einer älteren Version von Word 95 (Betriebssystem: Windows 95). Sie speichern Ihre Daten auf Diskette.

- Ihr neuer Rechner, den Sie sich angeschafft haben, läuft unter Windows XP mit Word 2003 und hat kein Diskettenlaufwerk mehr.
- Nun stehen Sie zunächst vor dem Problem, wie Sie Ihre Daten auf den neuen Rechner übertragen. Wenn Sie Glück haben, hat Ihr alter Rechner schon USB, so dass Sie Ihre Daten mit einem USB-Stick übertragen können. Vielleicht haben Sie auch noch ein Diskettenlaufwerk, auf das Sie zurückgreifen können. Oder aber Ihr alter Rechner kann sich ins Internet einwählen und Ihre Daten können von dort mit dem neuen Rechner runtergeladen werden. Hier ist unter Umständen ein wenig zu tun, es gibt jedoch noch genügend Möglichkeiten, Ihre Daten zu übertragen.
- Nehmen wir an, Ihre Daten sind sicher und korrekt übertragen worden. Wenn Sie Glück haben, meldet sich Word 2003 und sagt, Ihre Dateien seien in einem alten .doc-Format gespeichert und müssen in das aktuelle Format konvertiert werden. Diese Konvertierung ist dann eine Migration in ein neues, aktuelleres .doc-Format. Wenn die Migration erfolgreich abläuft, sieht Ihr Dokument aus wie auf dem alten Rechner unter Word 95, es besteht jedoch die Möglichkeit, dass Ihr Dokument sich verändert hat (Formatierung, Schriftart, Schriftgrösse, etc.).
- Sollten Sie Pech haben, erkennt Word das alte Format nicht und eine Migration ist nicht automatisch möglich. Dann bleibt noch die Möglichkeit, die alten Dateien mit einem Zwischenschritt über ein anderes Textformat, das beide Textprogramme beherrschen, zu konvertieren. Sicherlich können beide Programme einfache Textdateien verarbeiten (.txt), vielleicht auch Dateien im Rich-Text-Format (.rtf). Sie müssen nun Ihre Dokumente mit dem alten Word alle als Text- oder RTF-Datei neu speichern, diese erneut (wie oben beschrieben) auf den neuen Rechner übertragen und dann mit dem neuen Word (als Text- oder RTF-Datei) wieder öffnen. Sehr wahrscheinlich sind dann sehr viele Formatierungen (Inhaltsverzeichnisse, Überschriften, Schriftdicken, Schriftarten, etc.) verlorengegangen, da eine .txt-Datei keinerlei solcher Dinge speichern kann, nur der Text entspricht dem originalen Dokument. Mit einer RTF-Datei haben Sie sicherlich weniger Informationsverlust. Sie führen also praktisch zwei Migrationen durch: .doc (Word 95) ---> .txt (bzw. .rtf) ---> .doc (Word 2003), siehe hierzu die Abbildungen 12.2.1 und 12.2.2.

Beispiel:Zeichenkodierungen

- Eine Organisation, die in den 80er Jahren ihre Daten mit IBM Mainframes bearbeitet hat, möchte diese Daten auch auf späteren Systemen nutzen können. Die IBM Mainframes nutzten einen Zeichenstandard

kopal

Pressemitteilung

Test des für kopal entwickelten Archivsystems erfolgreich
Fazit der Testphase zur Abnahme durch die Projektpartner

Das kopal-System für die Langzeitarchivierung digitaler Objekte hat am 28.11.2005 eine erste 30-tägige Testphase erfolgreich durchlaufen. Damit wurde das von der IBM Deutschland für das Projekt kopal weiterentwickelte System DIAS (Digital Information Archiving System) durch die Projektpartner Die Deutsche Bibliothek, die Niedersächsische Staats- und Universitätsbibliothek (SUB) Göttingen und die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) abgenommen. Die Deutsche Bibliothek und die SUB Göttingen haben eine Funktionsprüfung des DIAS im Fernzugriff auf das bei der GWDG gehostete System durchgeführt. Beide Bibliotheken haben Testobjekte aus ihren elektronischen Beständen in die Tests eingebracht wie etwa Dissertationen, Zeitschriften und Monographien sowie Objekte des Göttinger Digitalisierungszentrums. Das wissenschaftliche Rechenzentrum GWDG stellte die Testumgebung zur Verfügung und führte Tests bezüglich betriebsüblichen Störungen und Systemausfällen durch.

Die von Mitte Oktober bis Ende November durchgeführten Tests schließen die beiden ersten Arbeitspakete der Entwicklungsphase des Projektes ab. Die Weiterentwicklung des DIAS zu einem kooperativ betriebenen Archivsystem wird unter Anwendung internationaler Standards von IBM übernommen. Die Projektmitarbeiter Der Deutschen Bibliothek und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen erstellen auf den weiterentwickelten DIAS-Core abgestimmte Open Source Softwareprodukte. Diese kopal-Tools ermöglichen eine einfache und skalierbare Einspeisung der Daten in das Archiv und einen komfortablen Zugriff auf die archivierten Materialien.

In der nun abgeschlossenen Testphase wurde die Funktionalität des DIAS im Zusammenspiel mit den entwickelten Tools zum automatisierten Einspeisen (Ingest) und Abfragen (Access) untersucht. Dokumentiert wurde außerdem das Verhalten des Systems bezüglich der Größe der Objekte sowie Vollständigkeit und Korrektheit der Lieferungen. Defekte oder unvollständige Datenpakete wurden eingespielt, um die im System integrierte Fehlererkennung zu testen. Es wurden mögliche Störungen und Systemausfälle simuliert, die erfahrungsgemäß im praktischen Betrieb vorkommen.

Für die produktive Nutzung der kopal-Lösung ist es erforderlich, dass verschiedene Mandanten ein hinreichend performantes System vorfinden, das in der Lage ist, Objekte zu handhaben, die von beiden Institutionen archiviert werden. Die Entwicklung von Komponenten für Administration und Error-Tracking befindet sich in der Vorbereitung.

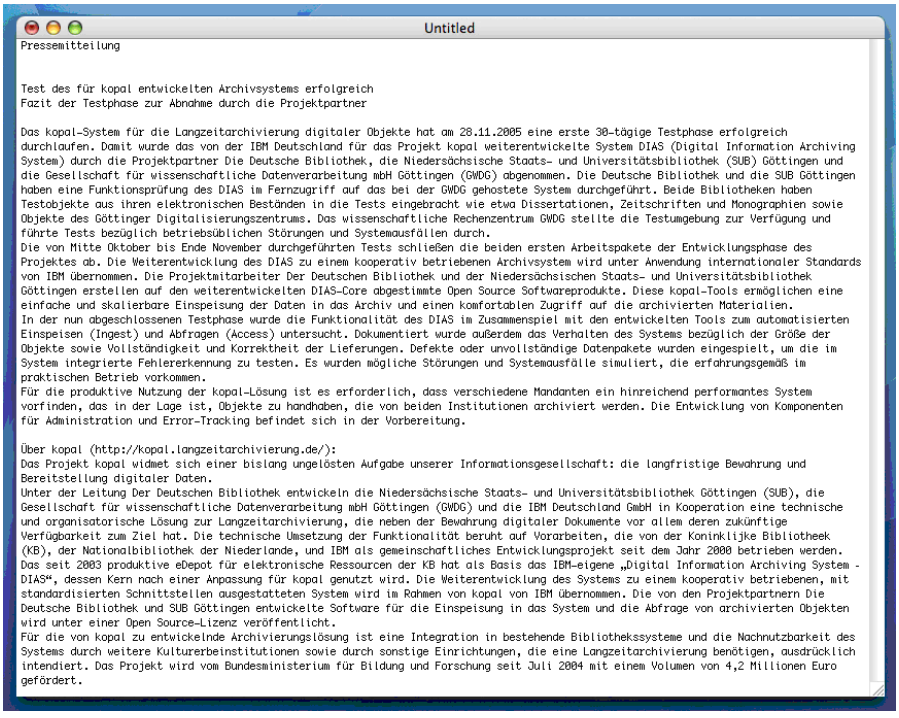
Über kopal (http://kopal.langzeitarchivierung.de/):

Das Projekt kopal widmet sich einer bislang ungelösten Aufgabe unserer Informationsgesellschaft: die langfristige Bewahrung und Bereitstellung digitaler Daten.

Unter der Leitung Der Deutschen Bibliothek entwickeln die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB), die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und die IBM Deutschland GmbH in Kooperation eine technische und organisatorische Lösung zur Langzeitarchivierung, die neben der Bewahrung digitaler Dokumente vor allem deren zukünftige Verfügbarkeit zum Ziel hat. Die technische Umsetzung der Funktionalität beruht auf Vorarbeiten, die von der Koninklijke Bibliotheek (KB), der Nationalbibliothek der Niederlande, und IBM als gemeinschaftliches Entwicklungsprojekt seit dem Jahr 2000 betrieben werden. Das seit 2003 produktive eDepot für elektronische Ressourcen der KB hat als Basis das IBM-eigene „Digital Information Archiving System - DIAS", dessen Kern nach einer Anpassung für kopal genutzt wird. Die Weiterentwicklung des Systems zu einem kooperativ betriebenen, mit standardisierten Schnittstellen ausgestatteten System wird im Rahmen von kopal von IBM übernommen. Die von den Projektpartnern Die Deutsche Bibliothek und SUB Göttingen entwickelte Software für die Einspeisung in das System und die Abfrage von archivierten Objekten wird unter einer Open Source-Lizenz veröffentlicht.

Für die von kopal zu entwickelnde Archivierungslösung ist eine Integration in bestehende Bibliothekssysteme und die Nachnutzbarkeit des Systems durch weitere Kulturerbeinstitutionen sowie durch sonstige Einrichtungen, die eine Langzeitarchivierung benötigen, ausdrücklich intendiert. Das Projekt wird vom Bundesministerium für Bildung und Forschung seit Juli 2004 mit einem Volumen von 4,2 Millionen Euro gefördert.

Abbildung 12.2.1: Ein Word-Dokument mit Grafiken, Formatierungen, Link, etc.



Untitled

Pressemitteilung

Test des für kopal entwickelten Archivsystems erfolgreich
Fazit der Testphase zur Abnahme durch die Projektpartner

Das kopal-System für die Langzeitarchivierung digitaler Objekte hat am 28.11.2005 eine erste 30-tägige Testphase erfolgreich durchlaufen. Damit wurde das von der IBM Deutschland für das Projekt kopal weiterentwickelte System DIAS (Digital Information Archiving System) durch die Projektpartner Die Deutsche Bibliothek, die Niedersächsische Staats- und Universitätsbibliothek (SUB) Göttingen und die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) abgenommen. Die Deutsche Bibliothek und die SUB Göttingen haben eine Funktionsprüfung des DIAS im Fernzugriff auf das bei der GWDG gehostete System durchgeführt. Beide Bibliotheken haben Testobjekte aus ihren elektronischen Beständen in die Tests eingebracht wie etwa Dissertationen, Zeitschriften und Monographien sowie Objekte des Göttinger Digitalisierungszentrums. Das wissenschaftliche Rechenzentrum GWDG stellte die Testumgebung zur Verfügung und führte Tests bezüglich betriebsüblichen Störungen und Systemausfällen durch.
Die von Mitte Oktober bis Ende November durchgeführten Tests schließen die beiden ersten Arbeitspakete der Entwicklungsphase des Projektes ab. Die Weiterentwicklung des DIAS zu einem kooperativ betriebenen Archivsystem wird unter Anwendung internationaler Standards von IBM übernommen. Die Projektmitarbeiter Der Deutschen Bibliothek und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen erstellen auf den weiterentwickelten DIAS-Core abgestimmte Open Source Softwareprodukte. Diese kopal-Tools ermöglichen eine einfache und skalierbare Einspeisung der Daten in das Archiv und einen komfortablen Zugriff auf die archivierten Materialien.
In der nun abgeschlossenen Testphase wurde die Funktionalität des DIAS im Zusammenspiel mit den entwickelten Tools zum automatisierten Einspeisen (Ingest) und Abfragen (Access) untersucht. Dokumentiert wurde außerdem das Verhalten des Systems bezüglich der Größe der Objekte sowie Vollständigkeit und Korrektheit der Lieferungen. Defekte oder unvollständige Datenpakete wurden eingespielt, um die im System integrierte Fehlererkennung zu testen. Es wurden mögliche Störungen und Systemausfälle simuliert, die erfahrungsgemäß im praktischen Betrieb vorkommen.
Für die produktive Nutzung der kopal-Lösung ist es erforderlich, dass verschiedene Mandanten ein hinreichend performantes System vorfinden, das in der Lage ist, Objekte zu handhaben, die von beiden Institutionen archiviert werden. Die Entwicklung von Komponenten für Administration und Error-Tracking befindet sich in der Vorbereitung.

Über kopal (http://kopal.langzeitarchivierung.de/):
Das Projekt kopal widmet sich einer bislang ungelösten Aufgabe unserer Informationsgesellschaft: die langfristige Bewahrung und Bereitstellung digitaler Daten.
Unter der Leitung Der Deutschen Bibliothek entwickeln die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB), die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und die IBM Deutschland GmbH in Kooperation eine technische und organisatorische Lösung zur Langzeitarchivierung, die neben der Bewahrung digitaler Dokumente vor allem deren zukünftige Verfügbarkeit zum Ziel hat. Die technische Umsetzung der Funktionalität beruht auf Vorarbeiten, die von der Koninklijke Bibliotheek (KB), der Nationalbibliothek der Niederlande, und IBM als gemeinschaftliches Entwicklungsprojekt seit dem Jahr 2000 betrieben werden. Das seit 2003 produktive eDepot für elektronische Ressourcen der KB hat als Basis das IBM-eigene „Digital Information Archiving System - DIAS", dessen Kern nach einer Anpassung für kopal genutzt wird. Die Weiterentwicklung des Systems zu einem kooperativ betriebenen, mit standardisierten Schnittstellen ausgestatteten System wird im Rahmen von kopal von IBM übernommen. Die von den Projektpartnern Die Deutsche Bibliothek und SUB Göttingen entwickelte Software für die Einspeisung in das System und die Abfrage von archivierten Objekten wird unter einer Open Source-Lizenz veröffentlicht.
Für die von kopal zu entwickelnde Archivierungslösung ist eine Integration in bestehende Bibliothekssysteme und die Nachnutzbarkeit des Systems durch weitere Kulturerbeinstitutionen sowie durch sonstige Einrichtungen, die eine Langzeitarchivierung benötigen, ausdrücklich intendiert. Das Projekt wird vom Bundesministerium für Bildung und Forschung seit Juli 2004 mit einem Volumen von 4,2 Millionen Euro gefördert.

Abbildung 12.2.2: Das selbe Dokument im .txt-Format ohne Formatierungen

namens EBCDIC[7].

- In den 90er Jahren installierte Rechner nutzten den ASCII Zeichencode (American National Standard Code for Information Interchange), welcher nicht alle Zeichen des EBCDIC abdeckte. Die Organisation mußte sich nun entscheiden, ob sie alle Dokumente nach ASCII konvertierten (und einen permanenten Verlust von Daten hinnahmen) oder sie nur bei Bedarf in ASCII umwandelten und die Originaldaten in EBCDIC beließen. So hatte man den gleichen Verlust beim Umwandeln, jedoch für spätere Zeit die Originaldaten erhalten.
- Bei Jahrtausendwechsel begann UNICODE[8] die Welt zu erobern und tatsächlich enthält UNICODE alle Zeichen des EBCDIC, so dass nun alle Dokumente 1:1 von EBCDIC in UNICODE konvertiert werden konnten (sofern die Originaldateien noch existierten!). Bei einer sofortigen Konvertierung in ASCII wären tatsächlich Daten verloren gegangen.

Zusammenfassung: Vor- und Nachteile von Migration

Vorteile von Migration

- Migration ist technisch (verglichen mit Emulation) gut zu realisieren.
- Migration kann in vielen Fällen automatisiert werden.
- Die migrierten Dokumente sind unabhängig von weiteren Komponenten (abgesehen von der aktuellen Darstellungssoftware).

- Die originalen Objekte können aufbewahrt werden, um evtl. später darauf zurückgreifen zu können.

Nachteile von Migration

- Jedes Objekt muss einzeln migriert werden.
- Die Wahrscheinlichkeit von Datenverlust bzw. Datenveränderung ist (besonders über mehrere Migrationsschritte) sehr hoch.
- Jede Version (Migration) eines Objekts inclusive des Original-Dokuments sollte gespeichert werden. Damit ist unter Umständen ein hoher Speicherplatzbedarf verbunden.
- Für jedes Format und für jeden Migrations-Schritt muss es ein Migrations-Werkzeug geben.

---

7 Extended Binary Coded Decimal Interchange Code, siehe <http://www.natural-innovations.com/computing/asciiebcdic.html>

8 Siehe <http://www.unicode.org>

- Migration ist nicht für alle Formate realisierbar.

Literatur

- CCSDS: Reference Model for an Open Archival Information System (OAIS) (2002) <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf> (letzter Zugriff: 7. Juni 2006)
- Jenkins, Clare: Cedars Guide to: Digital Preservation Strategies (2002) <http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html> (letzter Zugriff: 7. Juni 2006)

12.3 Emulation

Stefan E. Funk

Mit Emulation (Nachbildung, Nachahmung, von lat. aemulator = Nacheiferer) versucht man die auftretenden Verluste einer Datenformatmigration zu umgehen, indem man die originale Umgebung der archivierten digitalen Objekte nachbildet. Emulation kann auf verschiedenen Ebenen stattfinden:

- Zum einen auf der Ebene von Anwendungs-Software,
- zum anderen auf der Ebene von Betriebssystemen und zu guter Letzt
- auf der Ebene von Hardware-Plattformen.

So kann zum Beispiel die originale Hardware des digitalen Objekts als Software mit einem Programm nachgebildet werden, welches das archivierte Betriebssystem und die darauf aufbauenden Softwarekomponenten laden kann (Emulation von Hardware-Plattformen). Ein Beispiel für die Emulation von Betriebssystemen wäre ein MS-DOS-Emulator[9], der die Programme für dieses schon etwas ältere Betriebssystem auf aktuellen Rechnern ausführen kann. Ein Beispiel für den ersten Fall wäre etwa ein Programm zum Anzeigen und Bearbeiten von sehr alten Microsoft Word-Dateien (.doc), die das aktuelle Word nicht mehr lesen kann. Auf diese Weise wird die Funktionalität dieser alten und nicht mehr verfügbaren Soft- oder Hardware emuliert und die Inhalte bzw. die Funktionalität der damit erstellten Dokumente erhalten.

Im Gegensatz zur Migration, bei der jeweils eine neue und aktuellere Version des digitalen Objektes selbst erzeugt wird, werden die originalen Objekte bei der Emulation nicht verändert. Stattdessen muss man für jede neue Hardwarearchitektur die Emulationssoftware anpassen, im schlechtesten Fall muss diese jedes Mal neu entwickelt werden. Wenn das aus irgendeinem Grund nicht geschieht, ist der komplette Datenbestand der betroffenen Objekte unter Umständen nicht mehr nutzbar und damit für die Nachwelt verloren.

Emulation von Anwendungssoftware

Da es um die Darstellung der digitalen Dokumente geht, die wir vorhin beschrieben haben, ist die Emulation der Software, die mit diesen Dokumenten

9 DOS - Disc Operating System, näheres unter <http://www.operating-system.org/betriebssystem/_german/bs-msdos.htm>

arbeitet, eine erste Möglichkeit. So kann auf einem aktuellen System ein Programm entwickelt werden, das archivierte digitale Objekte in einem bestimmten Format öffnen, anzeigen oder bearbeiten kann, auf die mit aktueller Software auf diesem System nicht mehr zugegriffen werden kann, weil vielleicht die Original-Software nicht mehr existiert oder auf aktuellen Systemen nicht mehr lauffähig ist.

Wenn wir zum Beispiel eine PDF-Datei aus dem Jahr 1998, Version 1.2, darstellen möchten, und der aktuelle Acrobat Reader 7.0 stellt das Dokument nicht mehr richtig dar, müssen wir einen PDF-Reader für diese PDF-Version auf einem aktuellen Betriebssystem programmieren, sprich: einen alten PDF-Reader emulieren. Dieser sollte dann alle PDF-Dateien der Version 1.2 darstellen können. Für jeden Generationswechsel von Hardware oder Betriebssystem würde so ein PDF-Reader benötigt, um den Zugriff auf die PDF-Dokumente in Version 1.2 auch in Zukunft zu gewährleisten. Die genaue Kenntnis des PDF-Formats ist hierzu zwingend erforderlich.

Emulation von Betriebssystemen und Hardware-Plattformen

Bei einigen Anwendungen kann es sinnvoll sein, eine komplette Hardware-Plattform zu emulieren, zum Beispiel wenn es kein einheitliches Format für bestimmte Anwendungen gibt. Hier ist der Commodore-64 ein gutes Beispiel. Die Spiele für den C-64 waren eigenständige Programme, die direkt auf dem Rechner liefen, soll heissen, es wird direkt die Hardware inklusive des Betriebssystems[10] benötigt und nicht ein Programm, das diese Spiele ausführt (wie ein PDF-Viewer).

Es muss also ein Commodore-64 in Software implementiert werden, der sich genau so verhält wie die Hardware und das Betriebssystem des originalen Commodore-64 und auf einem aktuellen Computersystem lauffähig ist. Diese C-64-Emulatoren gibt es für nahezu alle aktuellen Computersysteme und auch weitere Emulatoren für andere ältere Systeme sind erhältlich[11].

10 Eine Trennung von Hardware und Betriebssystem ist beim Commodore-64 nicht nötig, da diese beiden Komponenten sehr eng zusammenhängen. Auch andere „Betriebssysteme“ wie zum Beispiel GEOS setzen auf das Betriebssystem des C-64 auf.

11 Hier einige Adressen im Internet zum Thema Emulatoren: <http://www.luke-web.de/games/emu.html>, <http://www.aep-emu.de/Emus.html>, <http://www.homecomputermuseum.de/>

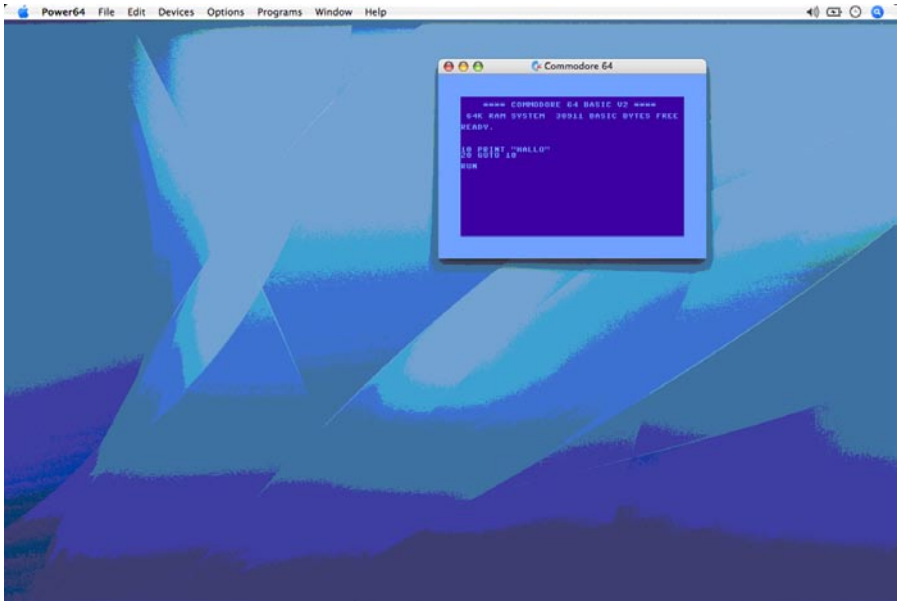


Abbildung 12.3.1: Power 64, ein Commodore-64 Emulator für Mac OS X

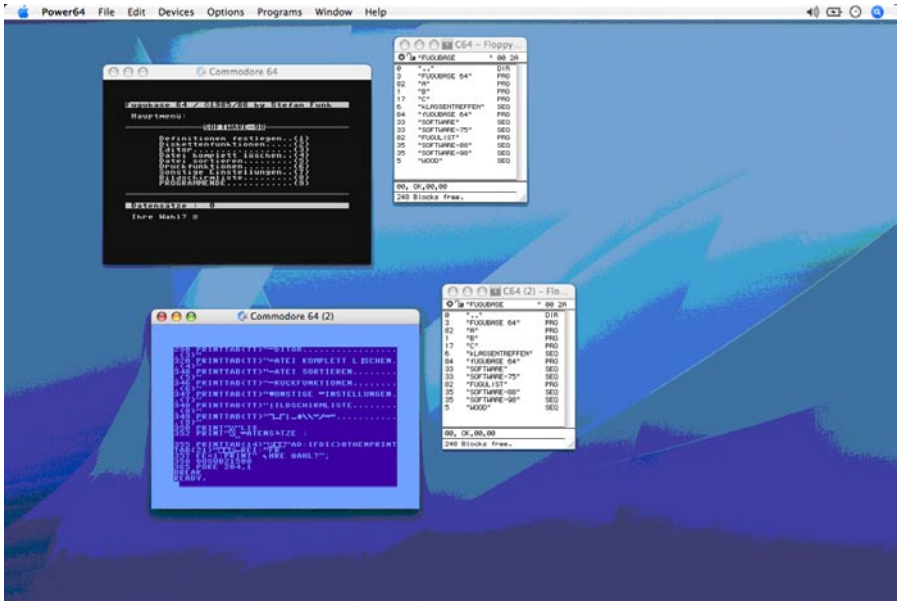


Abbildung 12.3.2: Fugubase 64, ein Datenverwaltungs-Programm in Basic für den C-64, emuliert unter Mac OS X (S. E. Funk, 1985/86)

Die Emulation eines Betriebssystems oder einer Hardware-Plattform ist eine sehr komplexe Sache, die schon für einen C-64-Emulator sehr viel Arbeit bedeutet. Man kann jedoch auch die Hardware eines PC in Software nachbilden, um dann auf einem solchen virtuellen PC beliebige Betriebssysteme und die auf ihnen laufenden Anwendungsprogramme oder auch Spiele zu starten (die Betriebssysteme wie auch die Programme bleiben dann im Originalzustand). Dies bedeutet im Allgemeinen, dass eine gute Performanz auf der aktuellen Hardware vorhanden sein muss. Eine Emulation eines Commodore-64 auf einem aktuellen PC ist jedoch keine performanzkritische Anwendung. Für zukünftige Computersysteme, die unsere heutigen emulieren sollen, wird im Allgemeinen davon ausgegangen, dass deren Performanz weitaus höher ist als heute, sodass auch hier die Performanz für eine erfolgreiche Emulation ausreichen dürfte.

Beispiel: Migration und Emulation alter C-64 Programme

- Da der Commodore 64 ein sehr beliebter und weit verbreiteter Homecomputer war, gibt es sehr viele Emulatoren für nahezu alle aktuellen Computersysteme. Viele Videospiele, die es für den C-64 gab, sind im Internet als C-64 Disk-Image zu finden. Die darin enthaltenen Programme können dann mit den Emulatoren geladen und genutzt werden. Als alter C-64 Nutzer stand ich also nicht vor dem Problem, meine Spiele von alten 5,25-Zoll Disketten auf neuere Datenträger migrieren zu müssen. Ein Emulator für den Apple unter Mac OS X ist Power64[12], siehe Abbildung 12.3.1.
- Anders sah es hingegen für die Programme aus, die ich vor mehr als 20 Jahren auf dem C-64 selbst programmiert habe. Es handelt sich hier um viele Programme in Commodore-64 BASIC. Die Frage, die sich mir stellte, war nun die, ob und wie ich diese Daten von meinen alten (auf dem Original C-64 noch laufenden) 5,25 Zoll-Disketten von 1982 bis 1987 auf die Festplatte meines PC kopieren und ich diese Daten auch für den C-64-Emulator nutzen kann.
- Der erste Versuch, einfach ein vor einigen Jahren noch gebräuchliches 5,25 Zoll-Laufwerk[13] an den PC anzuschließen und die C-64 Daten am PC auszulesen, schlug zunächst einmal fehl. Grund hierfür waren die unterschiedlichen Dichten und die unterschiedlichen Dateisysteme der 5,25 Zoll-Disketten. Auf eine Diskette des C-64 war Platz für 170 KB,

12 <http://www.infinite-loop.at/Power64/index.html>

13 Den ersten Versuch unternahm ich vor etwa vier Jahren, 5,25-Zoll-Diskettenlaufwerke waren nicht mehr wirklich gebräuchlich, aber noch erhältlich. Heute werden selbst die 3,5-Zoll-Laufwerke schon nicht mehr mit einem neuen Rechner verkauft. Neue Medien zum Datenaustausch und zur Speicherung sind heute USB-Stick, DVD, CD-ROM und Festplatte.

damals einfache Dichte (single density). Die Disketten für den PC hatten jedoch doppelte Dichte (double density) oder gar hohe Dichte (high density), sodass das mit zur Verfügung stehende Diskettenlaufwerk die C-64 Disketten nicht lesen konnte.

- Nach kurzer Recherche entdeckte ich eine Seite im Internet (die Community für den C-64 ist immer noch enorm groß), die Schaltpläne für einige Kabel abbildete, mit denen man seinen PC mit den Diskettenlaufwerken seines C-64 verbinden konnte. Mit Hilfe des Programmes Star Commander[14], das unter DOS läuft, kann man damit seine Daten von C-64 Disketten auf seinen PC kopieren und auch gleich Disk-Images erstellen. Inzwischen kann man solche Kabel auch bestellen und muss nicht selbst zum Lötkolben greifen (Für die Nutzung dieses Programms muss natürlich eine lauffähige DOS-Version zur Verfügung stehen, ist keine verfügbar, kann evtl. eine emuliert werden :-)
- Nach diesen Aktionen kann ich nun meine alten selbst erstellten Programme auf vielen C-64 Emulatoren wieder nutzen, weiterentwickeln und spielen, wie in Abbildung 12.3.2 und 12.3.3 zu sehen ist (und das sogar auf mehreren virtuellen Commodore-64 gleichzeitig).

Beispiel: Eine Emulation in der Emulation

- Es ist nun auch möglich, einen Emulator wiederum zu emulieren, wenn ein weiterer Generationswechsel einer Hardwareplattform ansteht. Ein praktisches Beispiel ist ein Apple Notebook, das unter Mac OS X, einem Unix-basierten Betriebssystem, arbeitet. Auf diesem werden zwei Emulatoren und ein weiteres originales Betriebssystem gestartet.
- Auf diesem Rechner wird das Programm Q gestartet[15], das eine Hardware-Plattform emuliert (einen Pentium x86 mit diversen Grafik-, Sound- und weiteren Hardwarekomponenten). Es basiert auf dem CPU-Emulator QEMU[16]•
- Auf dieser virtuellen Hardwareplattform kann nun ein originales Windows 98 installiert werden, so dass man ein reguläres, altbekanntes Windows 98 auf diesem nicht-Windows-Rechner nutzen kann. Das installierte Windows 98 kann selbstverständlich alle Programme für Windows 98 ausführen, da es sich tatsächlich um ein originales Windows 98 handelt. Sogar ein Windows-Update über das Internet ist möglich.
- Jetzt kann natürlich auch ein C-64 Emulator für Windows, hier der

14 <http://sta.c64.org/sc.html>

15 <http://www.kberg.ch/q/>

16 <http://fabrice.bellard.free.fr/qemu/>

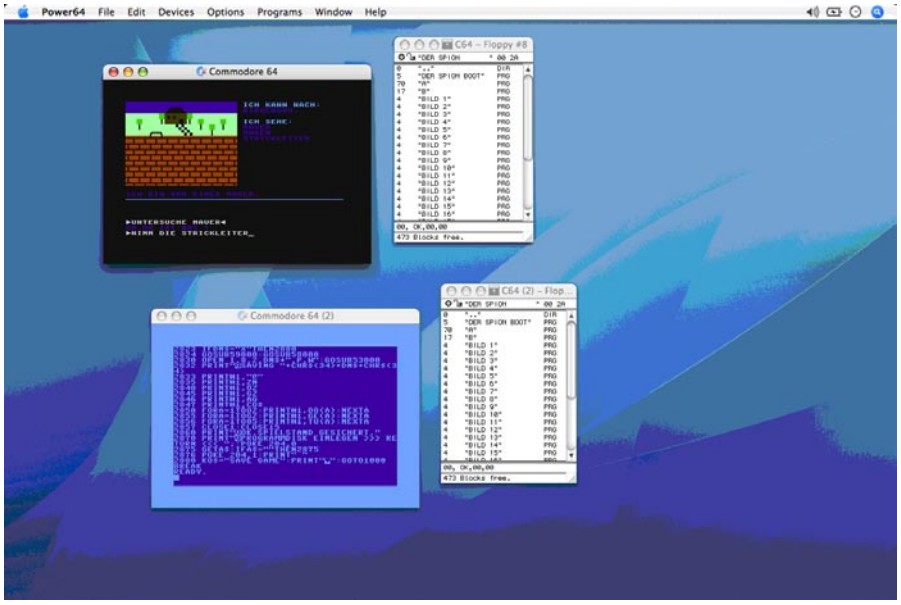


Abbildung 12.3.3: Der Spion, ein Adventure in Basic für den C-64, emuliert unter Max OS X (S. E. Funk, 1987)

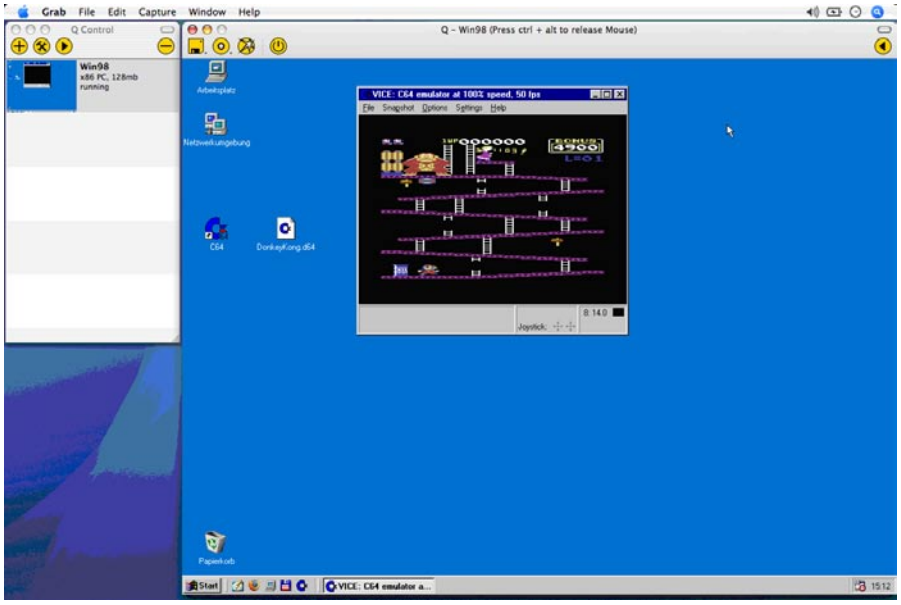


Abbildung 12.3.4: Das Videospiel Donkey Kong auf einem C-64 Emulator auf einem Windows 98 auf einem virtuellen Pentium auf einem Apple PowerBook unter Mac OS X

VICE[17], gestartet werden. Darauf laufen nun alle altbekannten und beliebten Commodore-64 Programme.

- Probleme kann es bei dieser Art von Emulation zum Beispiel bei der Performanz geben und je nach Qualität der Emulatoren auch mit hardwarespezifischen Dingen wie Grafik, Sound und angeschlossener Peripherie (Mäuse, Joysticks, etc.). Der C-64 Emulator muss schließlich durch Windows über die virtuelle Hardware (Emulation QEMU) auf die reale Hardware des Notebooks zugreifen. Bei steigender Komplexität solcher Emulationsszenarien wird die Anzahl der möglichen Fehler stark ansteigen. Als Beispiel siehe Abbildung 12.3.4.

Der Universal Virtual Computer (UVC)

Mittlerweile gibt es einen elaborierteren Ansatz der Emulation, den Universal Virtual Computer (UVC) von IBM. Der UVC ist ein wohldokumentierter virtueller Computer,der auf unterschiedlichen (auch zukünftigen) Architekturen nachgebildet werden kann. Aufgebaut ist er ähnlich wie heute existierende Computer, der beispielsweise Speicherzugriff ermöglicht. Mit Hilfe dieser Dokumentation ist es einem Programmierer auch auf zukünftigen Systemen möglich, diesen virtuellen Computer zu implementieren. Auf diesem virtuellen Computer aufbauend können nun Programme geschrieben werden, die zum Beispiel eine PDF-Datei lesen oder Grafiken darstellen können.

Archiviert wird jetzt der PDF-Reader (der Bildbetrachter), der für den UVC programmiert wurde, sowie das originale PDF-Dokument (oder die originale Grafik) selbst. Ein zukünftiger Nutzer kann dann auf einer zukünftigen und wahrscheinlich hoch entwickelten Hardware auch in ferner Zukunft noch mit Hilfe der Dokumentation des UVC einen solchen implementieren und mit Hilfe dieses virtuellen Computers den PDF-Reader starten, mit dem das archivierte PDF-Dokument dargestellt wird. Die Dokumentation muss selbstverständlich erhalten bleiben und lesbar sein.

Ein Problem dieser Idee ist sicherlich, dass bei zunehmendem Anspruch an die Emulation, die auf dem UVC laufen soll, eine Programmierung derselben immer schwieriger wird. Es wird sehr kompliziert, wenn für den UVC ein Betriebssystem wie Linux oder Windows programmiert werden soll, mit dessen Hilfe dann die Applikationen von Linux oder Windows genutzt werden können. Schon eine nachprogrammierte Version eines Textverarbeitungsprogrammes

17 http://www.viceteam.org/

wie zum Beispiel Word, mit dem später alte Word-Dokumente (.doc) auf dem UVC gelesen und bearbeitet werden können, ist ein höchst umfangreiches Unternehmen. Zumal hier nicht nur die Formatbeschreibung, sondern auch alle Programmfunktionen bekannt sein müssen.

Zusammenfassung: Vor- und Nachteile von Emulation

Vorteile von Emulation

- Bei der Emulation bleiben die Originalobjekte unverändert.
- Eine Konvertierung der Objekte ist nicht nötig.
- Für die Emulation wird weniger Speicherplatz benötigt, da keine Migrationen gespeichert werden müssen.

Nachteile von Emulation

- Für komplizierte Objekte/Systeme (wie Betriebssysteme oder Anwendungsprogramme) sind Emulatoren technisch schwer zu implementieren.
- Es entsteht ein hoher Aufwand pro Hardware-Generationswechsel. Es müssen für jede Plattform neue Emulatoren entwickelt werden.
- Die Spezifikationen für die zu emulierenden Objekte/Systeme sind nicht immer hinreichend bekannt.

Literatur

- Lorie, Raymond: the UVC: a method for preserving digital documents - proof of concept (2002) <http://www-5.ibm.com/nl/dias/resource/uvc.pdf> (letzter Zugriff: 4. Mai 2006)
- Nationaal Archief: Technical Description of the Universal Virtual Computer (UVC) - Data preservation process for spreadsheets (2005) <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/TDUVCv1.pdf> (letzter Zugriff: 6. Juni 2006)
- Erik Oltmans, Nanda Kol: A Comparison Between Migration and Emulation in Terms of Costs (2005) <http://www.rlg.org/en/page.php?Page_ID=20571#article0> (letzter Zugriff: 4. Mai 2006)

12.4 Computermuseum

Karsten Huth

Definition

Auch wenn man die Strategie der Hardware Preservation als Methode zur Langzeitarchivierung auf keinen Fall empfehlen sollte, so ist es leider alltägliche Praxis, dass digitale Langzeitarchive auch obsolete Hardware vorhalten müssen, zumindest bis sie in der Lage sind, besser geeignete Strategien durchzuführen. Aber gerade in den Anfängen eines digitalen Archivs, wenn es noch über keinen geregelten Workflow verfügt, werden digitale Objekte oft auf ihren originalen Datenträgern oder mitsamt ihrer originalen Hardware/Software Umgebung abgeliefert. Dies betrifft vor allem digitale Objekte, die technologisch obsolet geworden sind. Deshalb sind in der Praxis, wenn auch ungewollt, Computermuseen eher die Regel als eine Ausnahme.

Leider hat sich der Begriff „Computermuseum" im deutschen Sprachraum verfestigt. Passender wäre der Begriff „Hardware-/Software-Konservierung", denn die konservierten Computer müssen nicht unbedingt nur im Rahmen eines Museums erhalten werden. Man muss vielmehr differenzieren zwischen:

1. Hardware Preservation als Strategie zur Archivierung von digitalen Objekten:
   Eigentliches Ziel ist die Erhaltung der digitalen Objekte. Zu diesem Zweck versucht man die ursprüngliche Hardware/Software Plattform so lange wie möglich am Laufen zu halten.
2. Hardware Preservation im Rahmen eines Technikmuseums:
   Wird im ersten Fall die Hardware/Software Plattform nur erhalten, um den Zugriff auf die digitalen Objekte zu ermöglichen, so ist hier die ursprüngliche Hardware/Software Plattform das zentrale Objekt der konservatorischen Bemühungen. Während im ersten Fall Reparaturen an der Hardware einzig der Lauffähigkeit der Rechner dienen, so fallen im Rahmen eines Technikmuseums auch ethische Gesichtspunkte bei der Restauration ins Gewicht. Die Erhaltung der Funktion ist bei einer Reparatur nicht mehr das einzige Kriterium, es sollten auch möglichst die historisch adäquaten Bauteile verwendet werden. Diese Auflage erschwert die beinahe unmögliche Aufgabe der Hardware-Konservierung noch zusätzlich.

Bei einem technischen Museum liegt die Motivation zur Konservierung von Hardware auf der Hand. Die historische Hardware zusammen mit der originalen Software sind die Sammelobjekte und Exponate des Museums. Deswe-

gen müssen sie solange wie möglich ein einem präsentablen Zustand erhalten werden. Daneben gibt es aber auch noch weitere Gründe, die für die Hardware Preservation als Archivierungsstrategie sprechen.

Gründe zur Aufrechterhaltung eines Computermuseums:

- Keine andere Strategie erhält soviel vom intrinsischen Wert der digitalen Objekte (Look and Feel). An Authentizität ist dieser Ansatz nicht zu übertreffen.[18]
- Bei komplexen digitalen Objekten, für die Migration nicht in Frage kommt, und eine Emulation der Hardware/Software Umgebung noch nicht möglich ist, ist die Hardware Preservation die einzige Möglichkeit, um das Objekt zumindest für einen Übergangszeitraum zu erhalten.[19]
- Zur Unterstützung von anderen Archivierungsstrategien kann die zeitweise Erhaltung der originalen Plattformen notwendig sein. Man kann z. B. nur durch einen Vergleich mit der ursprünglichen Hardware/Software Plattform überprüfen, ob ein Emulatorprogramm korrekt arbeitet oder nicht.[20]

Probleme der Hardware Preservation:

Ob man ein Hardware-Museum aus dem ersten oder dem zweiten Grund führt, in beiden Fällen hat man mit den gleichen Problemen zu kämpfen. Zum einen ergeben sich auf lange Sicht gesehen große organisatorische und zum anderen rein technische Probleme der Konservierung von Hardware und Datenträgern.

1. Organisatorische Probleme:

- Die Menge an zu lagerndem und zu verwaltendem Material wird stetig wachsen. Da nicht nur die Rechner sondern auch Peripheriegeräte und Datenträger gelagert werden müssen, steigt der Platzbedarf und der Lagerungsaufwand enorm an. „Selbst heute schon erscheint es unrealistisch, sämtliche bisher entwickelten Computertypen in einem Museum zu ver-

18 Borghoff, Uwe M. et al. (2003): *Methoden zur Erhaltung digitaler Dokumente*. 1. Aufl. Heidelberg : dpunkt-Verl., 2003: S. 16-18

19 Jones, Maggie/ Beagrie, Neil (o.J.): *Preservation Management of Digital Materials: A Handbook*. Digital Preservation Coalition. < http://www.dpconline.org/text/orgact/storage.html> (Abrufdatum: 14.12.2007)

20 Rothenberg, Jeff (1998): *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources*. Washington D.C.: Council on Library and Information Resources: S. 12-13 <http://www.clir.org/pubs/reports/rothenberg/inadequacy.html> (Abrufdatum 14.12.2007)

sammeln, geschweige denn dies für die Zukunft sicher zu stellen."[21]

- Techniker und Experten, die historische Computer bedienen und gegebenenfalls reparieren können, werden über kurz oder lang nicht mehr zur Verfügung stehen. Mit wachsendem Bestand müssten die Mitarbeiter des Museums ihr Fachwissen ständig erweitern, oder der Bedarf an Technikexperten und neuen Mitarbeitern würde ständig wachsen.[22]
- Die Nutzung der digitalen Objekte ist nur sehr eingeschränkt möglich. Da die obsoleten Computersysteme von der aktuellen Technologie abgeschnitten sind, könnte der Nutzer nur im Computermuseum auf die Objekte zugreifen.[23]

2. Technische Probleme:

- Die technischen Geräte und Bausteine haben nur eine begrenzte Lebenserwartung. Da für obsolete Systeme keine Ersatzteile mehr produziert werden, ist die Restaurierung eines Systems irgendwann nicht mehr möglich.[24]
- Neben der Hardware muss auch die originale Softwareumgebung erhalten und archiviert werden. Diese muss natürlich auf den entsprechenden Datenträgern vorgehalten werden. Da Datenträger ebenso wie die Hardware nur eine begrenzte Lebensdauer haben, müssen die Software und die Daten von Zeit zu Zeit auf neue, frischere Datenträger des gleichen Typs, oder zumindest auf passende Datenträger des gleichen Computersystems umkopiert werden. Da jedoch Datenträger eines obsoleten Systems nicht mehr hergestellt werden, stößt diese Praxis zwangsläufig an ihre Grenze, und Software und Daten gehen verloren.[25]

Auftretende Schäden bei der Lagerung:

Es gibt wenig Literatur über die tatsächlich in der Praxis auftretenden Schäden. Der folgende Abschnitt bezieht sich auf eine Umfrage in Computermuseen. Diese Umfrage war Teil einer Abschlussarbeit an der San Francisco State University im Fach Museum Studies. Die folgende Aufzählung ist eine vorläufige Rangliste der auftretenden Probleme.[26]

21 s. Borghoff (2003)

22 Dooijes, Edo Hans (200): *Old computers, now and in the future.* Department of Computerscience/University of Amsterdam. <http://www.science.uva.nl/museum/pdfs/oldcomputers_dec2000.pdf> (Abrufdatum: 14.12.2007)

23 s. Rothenberg (1998)

24 s. Borghoff (2003)

25 s. Rothenberg (1998)

26 Gibson, Mark A. (2006): *The conservation of computers and other high-tech artifacts . Unique problemes and long-term solutions*: Thesis M.A. San Francisco : San Francisco State University

- Zerfall von Gummiteilen: Gummi wird für viele Bauteile der Hardware verwendet. Riemen in Motoren, Rollen in Magnetbänderlaufwerken, Lochkartenleser und Drucker, um nur einige Beispiele zu nennen. Gummi ist anfällig für Oxidation. Harte Oberflächen werden durch Oxidation weich und klebrig. Mit fortschreitendem Zerfall kann der Gummi wieder verhärten und dabei brüchig werden.
- Zerfall von Schaumstoffisolierungen: Schaumstoff wird hauptsächlich zur Lärmisolierung und Luftfilterung in Computern verwendet. Vor allem Schaumstoff aus Polyurethan ist sehr anfällig für eine ungewollte Oxidation. Das Material verfärbt sich zunächst und zerfällt dann in einzelne Krümel.
- Verfärbung von Plastikteilen: UV-Licht verändert die chemische Zusammensetzung der Plastikgehäuse. Die Funktion des Geräts wird dadurch zwar nicht beeinträchtigt, aber die Farbe des Gehäuses verändert sich merklich ins Gelb-bräunliche.
- Schäden durch Staub: Staub greift sowohl das Äußere der Hardware als auch ihr Innenleben an. Staub ist nur eine grobe Umschreibung für eine Vielzahl an Schadstoffen, wie z.B. Ruß, Ammoniumnitrat, Ammoniumsulfat und Schwefelsäure. Mit dem Staub lagert sich Salz und Feuchtigkeit an den Bauteilen ab. Dadurch wird die Anfälligkeit für Rost oder Schimmel erhöht. Lüfter mit Ventilatoren zur Kühlung von Prozessoren ziehen den Staub in das Gehäuse des Rechners.
- Zerfall der Batterien: Leckende Batterien können das Innenleben eines Rechners zerstören. Batterien sind Behälter bestehend aus Metal und Metaloxid eingetaucht in eine Flüssigkeit oder ein Gel aus Elektrolyten. Batterien sind sehr anfällig für Rost. Bei extrem unsachgemäßer Behandlung können sie sogar explodieren. Austretende Elektrolyte können Schaltkreise zersetzen.
- Rost: Metall ist ein häufiger Werkstoff in elektronischen Geräten. Anfällig für Rost sind Eisen, Stahl und Aluminium. Metall wird vor allem für das Gehäuse sowie für Klammern, Schrauben und Federn verwendet.
- Beschädigte Kondensatoren: Ähnlich wie bei einer Batterie ist ein Elektrolyt wesentlicher Bestandteil eines Kondensators. Das Elektrolyt kann eine Flüssigkeit, eine Paste oder ein Gel sein. Problematisch wird es, wenn das Elektrolyt austrocknet, da dann der Kondensator nicht mehr arbeitet. Trocknet das Elektrolyt nicht aus, kann der Kondensator lecken, so dass das Elektrolyt austritt, und ähnlichen Schaden anrichtet, wie eine kaputte Batterie. Kondensatoren die lange ungenutzt bleiben können explodieren.
- Zerfall des Plastiks: Plastik löst sich über einen längeren Zeitraum hin-

weg auf. Der sogenannte Weichmacher, ein chemischer Stoff, der bei der Produktion beigemengt wird, tritt in milchartigen Tropfen aus dem Material aus. Bei bestimmten Plastiksorten riecht die austretende Feuchtigkeit nach Essig. Der Prozess beeinträchtigt auch die Haltbarkeit von anderen Materialien, die mit dem zerfallenden Plastik verbunden sind.
- Schimmel: Bei einigen Monitoren aus den siebziger und achtziger Jahren kann Schimmel an der Innenseite der Mattscheibe auftreten.

Stark gefährdete Geräte und Bauteile:

Von den oben genannten möglichen Schäden sind die folgenden Bauteile am häufigsten betroffen:
- Schaltkreise die auf Dauer ausfallen.
- Kondensatoren die ausfallen oder explodieren.
- Ausfall von Batteriebetriebenen Speicherkarten und EPROMS und ein damit einhergehender Datenverlust.
- Zerstörte Kartenleser und Magnetbandlaufwerke durch kaputte Gummirollen.
- Verstaubte und verschmutzte Kontakte.
- Gebrochene oder verlorengegangene Kabel.[27]

Gesundheitsschädliche Stoffe und Risiken

Zu beachten ist, dass Restauratoren mit gesundheitsgefährdenden Stoffen am Arbeitsplatz in Kontakt kommen können. Welche Stoffe in Frage kommen, hängt vom Alter und der Bauart der Hardware ab. Dokumentiert ist das Auftreten von:
- Quecksilber
- Blei (auch bleihaltige Farbe)
- Polychlorierten Biphenylen (PCB)
- Thorium u. anderen radioaktiven Substanzen
- Asbest
- Cadmium

Besondere Vorsicht ist beim Umgang mit Batterien (vor allem defekten, leckenden Batterien) und Kondensatoren geboten. Abgesehen davon, dass Kondensatoren oft gesundheitsgefährdende Stoffe enthalten, können sie auch in stillgelegtem Zustand über Jahre hin eine hohe elektrische Spannung aufrecht halten. Wenn Kondensatoren nach längerer Zeit wieder unter Strom gesetzt werden, können sie explodieren.[28]

27 s. Dooijes (2000)

28 s. Gibson (2006)

Empfehlung zur Lagerung und Restaurierung:

Die Hardware sollte bei der Lagerung möglichst vor Licht geschützt werden. Ideal ist ein Helligkeitswert um 50 Lux. Fensterscheiben sollten die UV-Strahlung herausfiltern. Dadurch wird der Zerfall von Plastik und Gummi verlangsamt. Ebenso ist eine möglichst niedrige Raumtemperatur, unter 20°C, sowie eine relative Luftfeuchtigkeit von unter 50% ratsam. Beides verlangsamt den Zerfall von Gummi und Plastik, die niedrige Luftfeuchtigkeit verringert die Wahrscheinlichkeit von Rost. Vor der Inbetriebnahme eines Rechners sollte abgelagerter Staub durch vorsichtiges Absaugen entfernt werden. Dabei ist erhöhte Sorgfalt geboten, damit keine elektrostatische Energie die Schaltkreise beschädigt und keine wichtigen Teile mit eingesaugt werden. Mit einer zuvor geerdeten Pinzette können gröbere Staubknäuel beseitigt werden. Batterien sollten während der Lagerung möglichst aus der Hardware entfernt werden. Weit verbreitete Batterietypen sollten nicht gelagert werden. Wenn die Hardware in Betrieb genommen wird, werden frische Batterien des betreffenden Typs eingesetzt. Seltene, obsolete Batterietypen sollten separat gelagert werden. Alle genannten Maßnahmen können den Zerfall der Hardware jedoch nur verlangsamen. Aufzuhalten ist er nicht. Defekte Bauteile werden oft durch das Ausschlachten von Hardware gleicher Bauart ersetzt. Dabei werden alle intakten Teile zu einer funktionierenden Hardwareeinheit zusammengefügt. Natürlich stößt dieses Verfahren irgendwann an seine Grenzen.
Bereits eingetretene Schäden sollten durch Restaurationsarbeiten abgemildert werden. Auslaufende Flüssigkeiten aus Kondensatoren oder Batterien sollte man umgehend mit Isopropanol-Lösung entfernen.

Dokumentation

Ein Computermuseum kommt natürlich um die korrekte Verzeichnung seiner Artefakte (Hardware und Software) nicht herum. Zusätzlich werden Informationen über den Betrieb, die Bedienung und die verwendete Technik der Hardware und Software benötigt. Des weiteren sollten Informationen über den Erhaltungszustand und potentiell anfällige Bauteile der Hardware erhoben und gesammelt werden. Wie bei anderen Erhaltungsstrategien fallen auch hier Metadaten an, die gespeichert und erschlossen werden wollen. Schon bei der Aufnahme eines obsoleten Systems in das Archiv sollte darauf geachtet werden, dass die notwendigen Zusatzinformation verfügbar sind (z.B. Betriebshandbücher über die Hardware/Software, technische Beschreibungen und Zeichnungen usw.). Da diese Informationen bei älteren Systemen meistens nur in gedruckter Form vorliegen, sollte auch hier Raum für die Lagerung mit einkalkuliert oder

eine Digitalisierung der Informationen erwogen werden.[29]

Beispieldaten des Computerspiele Museums Berlin

Die Softwaresammlung umfasst zurzeit 12.000 Titel über eine Zeitspanne von 1972 bis heute. Die Software wird getrennt von der Hardware in normalen Büroräumen gelagert und hat einen Platzbedarf von ca. 70 qm.
In der Hardwaresammlung des Computerspiele Museums befinden sich augenblicklich 2180 Sammlungsstücke. Sie sind in einer Datenbank inklusive Foto erfasst und inventarisiert. Die Sammlung besteht aus Videospielautomaten, Videospiele Konsolen, Heimcomputer, Handhelds, technische Zusatzteile (Laufwerke, Controller, Monitore etc.) Des weiteren besitzt das Museum eine umfangreiche Sammlung gedruckter Informationen wie Computerspiele Magazine und Handbücher. Diese sind in einer gesonderten Datenbank erfasst. Die Hardwaresammlung ist auf ca. 200 qm an der Peripherie Berlins untergebracht. Der Hauptgrund dafür ist, die günstigere Miete für die Räume als das in zentralerer Lage möglich wäre. Die Räume sind beheizbar und entsprechen größtenteils ebenfalls Bürostandard.[30]

29 s. Dooijes (2000)

30 Daten stammen von Herrn Andreas Lange, Kurator des Computerspielemuseums Berlin (2006)

12.5 Mikroverfilmung

Christian Keitel

Ein ungelöstes Problem bei der langfristigen Archivierung digitaler Informationen ist die begrenzte Haltbarkeit digitaler Datenträger. Künstliche Alterungstests sagen CDs, DVDs und Magnetbändern nur eine wenige Jahre währende Haltbarkeit voraus, während herkömmliche Trägermedien wie z.B. Pergament oder Papier mehrere Jahrhunderte als Datenspeicher dienen können. Hervorragende Ergebnisse erzielt bei diesen Tests insbesondere der Mikrofilm. Bei geeigneter (kühler) Lagerung wird ihm eine Haltbarkeit von über 500 Jahren vorausgesagt. Verschiedene Projekte versuchen daher, diese Eigenschaften auch für die Archivierung genuin digitaler Objekte einzusetzen. Neben der Haltbarkeit des Datenträgers sind dabei auch Aspekte wie Formate, Metadaten und Kosten zu bedenken.

In Anlehnung an die Sicherungs- und Ersatzverfilmung herkömmlicher Archivalien wurden zunächst digitale Informationen auf Mikrofilm als Bild ausbelichtet und eine spätere Benutzung in einem geeigneten Lesegerät (Mikrofilmreader) geplant. Erinnert sei in diesem Zusammenhang an das in den Anfängen des EDV-Einsatzes in Bibliotheken übliche COM-Verfahren (Computer Output on Microfilm/-fiche) zur Produktion von Katalog-Kopien. In letzter Zeit wird zunehmend von einer Benutzung im Computer gesprochen, was eine vorangehende Redigitalisierung voraussetzt. Dieses Szenario entwickelt die herkömmliche Verwendung des Mikrofilms weiter, sie mündet in einer gegenseitigen Verschränkung digitaler und analoger Techniken. Genuin digitale Daten werden dabei ebenso wie digitalisierte Daten von ursprünglich analogen Objekten/Archivalien auf Mikrofilm ausbelichtet und bei Bedarf zu einem späteren Zeitpunkt über eine spezielles Lesegerät redigitalisiert, um dann erneut digital im Computer benutzt zu werden. Eine derartige Konversionsstrategie erfordert im Vergleich mit der Verwendung des Mikrofilms als Benutzungsmedium einen wesentlich höheren Technikeinsatz.

Ein zweiter Vorteil liegt neben der Haltbarkeit des Datenträgers darin, dass die auf dem Mikrofilm als Bilder abgelegten Informationen nicht regelmäßig wie bei der Migrationsstrategie in neue Formate überführt werden müssen. Völlig unabhängig von Formaterwägungen ist der Mikrofilm jedoch nicht, da er über die Ablagestruktur von Primär- und v.a. Metadaten gewisse Ansprüche an das Zielformat bei der Redigitalisierung stellt, z.B. die bei den Metadaten ange-

wandte Form der Strukturierung. Die Vorteile im Bereich der Formate würden sich verlieren, wenn der Mikrofilm als digitales Speichermedium begriffen würde, um die Informationen nicht mehr als Bild, sondern als eine endlose Abfolge von Nullen und Einsen binär, d.h. als Bitstream, abzulegen.

Bei der Ausbelichtung der digitalen Objekte ist darauf zu achten, dass neben den Primärdaten auch die zugehörigen Metadaten auf dem Film abgelegt werden. Verglichen mit rein digitalen Erhaltungsstrategien kann dabei zum einen die für eine Verständnis unabdingaber Einheit von Meta- und Primärdaten leichter bewahrt werden. Zum anderen verspricht die Ablage auf Mikrofilm auch Vorteile beim Nachweis von Authentizät und Integrität, da die Daten selbst nur schwer manipuliert werden können (die Möglichkeit ergibt sich nur durch die erneute Herstellung eines Films).

Vor einer Abwägung der unterschiedlichen Erhaltungsstrategien sollten sowohl die Benutzungsbedingungen als auch die Kosten beachtet werden, die bei der Ausbelichtung, Lagerung und erneuten Redigitalisierung entstehen. Schließlich ist zu überlegen, in welcher Form die Informationen künftig verwendet werden sollen. Während der Einsatz des Mikrofilms bei Rasterbildern (nichtkodierten Informationen) naheliegt, müssen kodierte Informationen nach erfolgter Redigitalisierung erneut in Zeichen umgewandelt werden. Die Fehlerhäufigkeit der eingesetzten Software muss dabei gegen die zu erwartenden Vorteile aufgewogen werden.

Literatur

Projekt ARCHE, s. http://www.landesarchiv-bw.de >>> Aktuelles >>> Projekte

13 Access

Karsten Huth

Der Titel dieses Kapitels ist ein Begriff aus dem grundlegenden ISO Standard OAIS. Access steht dort für ein abstraktes Funktionsmodul (bestehend aus einer Menge von Einzelfunktionalitäten), welches im Wesentlichen den Zugriff auf die im Archiv vorgehaltenen Informationen regelt. Das Modul Access ist die Schnittstelle zwischen den OAIS-Modulen „Data Management", „Administration" und „Archival Storage".[1] Zudem ist das Access-Modul die Visitenkarte eines OAIS für die Außenwelt. Nutzer eines Langzeitarchivs treten ausschließlich über dieses Modul mit dem Archiv in Kontakt und erhalten gegebenenfalls Zugriff auf die Archivinformationen. In der digital vernetzten Welt kann man davon ausgehen, dass der Nutzer von zu Hause aus über ein Netzwerk in den Beständen eines Archivs recherchiert. Entsprechende technische Funktionalitäten wie Datenbankanfragen an Online-Kataloge oder elektronische Findmittel werden bei vielen Langzeitarchiven zum Service gehören. Die Möglichkeit von Fernanfragen an Datenbanken ist jedoch keine besondere Eigenart eines Langzeitarchivs. Wesentlich sind folgende Fragen:

- Wie können die Informationsobjekte (z. T. auch als konzeptuelle Objekte bezeichnet) dauerhaft korrekt adressiert und nachgewiesen werden,

1 Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information System: Blue Book*. Washington, DC. Page 4-14ff

wenn die logischen Objekte (z.B. Dateien, Datenobjekte) im Zuge von Migrationen technisch verändert werden und im Archiv in verschiedenen technischen Repräsentationen vorliegen?[2]

- Wie kann der Nutzer erkennen, dass die an ihn gelieferte Archivinformation auch integer und authentisch ist?[3]
- Wie kann das Archiv bei fortwährendem technologischem Wandel gewährleisten, dass die Nutzer die erhaltenen Informationen mit ihren verfügbaren technischen und intellektuellen Mitteln auch interpretieren können?

Erst wenn sich ein Archiv in Bezug auf den Zugriff mit den oben genannten Fragen befasst, handelt es strategisch im Sinne der Langzeitarchivierung. Die entsprechenden Maßnahmen bestehen natürlich zum Teil aus der Einführung und Implementierung von geeigneten technischen Infrastrukturen und Lösungen. Da die technischen Lösungen aber mit der Zeit auch veralten und ersetzt werden müssen, sind die organisatorischen, strategischen Maßnahmen eines Archivs von entscheidender Bedeutung. Unter diesem Gesichtspunkt sind Standardisierungen von globalen dauerhaften Identifikatoren, Zugriffsschnittstellen, Qualitätsmanagement und Zusammenschlüsse von Archiven unter gemeinsamen Zugriffsportalen eine wichtige Aufgabe für die nationale und internationale Gemeinde der Gedächtnisorganisationen.

2 vgl. Funk, Stefan: *Kap 9.1 Digitale Objekte*

3 nestor - Materialien 8: nestor - Kompetenznetzwerk Langzeitarchivierung / Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Version 1 (Entwurf zur Öffentlichen Kommentierung), Juni 2006, Frankfurt am Main : nestor c/o Die Deutsche Bibliothek, urn:nbn:de:0008-2006060710; Punkt 6.3 S. 16

13.1 Retrieval

Matthias Neubauer

Genauso wichtig wie die sichere Archivierung der digitalen Objekte ist auch die Möglichkeit, diese Objekte wieder aus dem Archiv herauszuholen und zu nutzen. Dabei muss gewährleistet sein, dass die Objekte den Zustand und den Informationsgehalt zum Zeitpunkt des Einspielens in das Archivsystem widerspiegeln. Im Idealfall sollte das Objekt noch exakt so abrufbar sein, wie es einmal in das Archiv eingespielt wurde. Je nach Verwendungszweck kann es jedoch auch sinnvoll sein, eher eine migrierte Form eines Objektes abzurufen. Einige wichtige Punkte, die es beim Zugriff von archivierten Objekten zu beachten gilt, sollen im Folgenden kurz erläutert werden.

Objektidentifikation

Zunächst ist eine eindeutige Identifikation des abzurufenden Objektes wichtig. Zu dieser Thematik existieren vielerlei Lösungen und Philosophien. Einige werden in den folgenden Kapiteln zum Thema „Persistent Identifier" vorgestellt. Grundsätzlich muss es anhand der verwendeten Identifizierungen möglich sein, jedwede Form und Version eines digitalen Objektes aus dem Langzeitarchiv abzurufen. Dies kann gegebenenfalls auch durch eine Kombination von externen und internen Identifikatoren realisiert werden.

Datenkonsistenz

Die Unversehrtheit der Daten hat höchste Priorität. Innerhalb des Archivs sollte durch geeignete Routinen zwar sichergestellt sein, dass der originale digitale Datenstrom erhalten bleibt. Jedoch können auch - und vor allem - bei der Übertragung der Daten aus dem Archiv heraus Inkonsistenzen durch Übertragungsfehler oder andere Störeinflüsse entstehen. Idealerweise sollte daher bei jedem Zugriff auf ein Archivobjekt über Checksummenvergleiche die Unversehrtheit der Daten sichergestellt werden. Je nach Art und Status der Daten kann diese Überprüfung auch nur stichprobenartig erfolgen.

Versionsmanagement

Je nach Verwendungszweck der Daten kann es entweder sinnvoll sein, das ursprüngliche Originalobjekt aus dem Archiv herauszuholen, oder

aber auch eine migrierte Form zu nutzen. Die höchste Authentizität wird man sicherlich mit dem ursprünglichen Objekt erreichen, jedoch kann es sich auf zukünftigen Systemen sehr schwierig gestalten, die erhaltenen Daten aufzubereiten und zu nutzen (mehr darüber im Kapitel über Emulation und Migration). Ein gutes Langzeitarchivierungssystem sollte nach Möglichkeit sowohl Originalversion und letzte Migrationsform, als auch alle dazwischen liegenden Objektversionen zugreifbar halten, um eine vollkommene Transparenz und Rekonstruierbarkeit zu gewährleisten.

Interpretation und Aufbereitung der Daten

Sofern das digitale Objekt zum Zweck einer Präsentation oder Weiternutzung abgerufen wurde, muss es durch geeignete Methoden aufbereitet und verfügbar gemacht werden. Schon beim Einspielen der Daten in das Archivsystem ist daher darauf zu achten, dass man die Struktur des Objektes in den beiliegenden Metadaten dokumentiert. Zudem kann es notwendig sein, die innerhalb eines Archivsystems verwendeten Schlüsselnummern zur eindeutigen Identifikation von Dateiformaten zu entschlüsseln und auf ein anderes System einzustellen.

Caching

Unter dem Begriff „Caching" versteht man die Pufferung oft genutzter Daten in einem schnell verfügbaren und hochperformanten Zwischenspeicher. Im Falle des Retrieval aus einem Langzeitarchivierungssystem ist dies dann sinnvoll, wenn die Archivobjekte auch als Basis für Präsentationssysteme und den täglichen Zugriff dienen sollen. Um das Archivsystem nicht mit unnötigen Anfragen nach häufig genutzten Objekten zu belasten, wird ein lokaler Zwischenspeicher angelegt, der stark frequentierte Objekte vorhält und gegebenenfalls mit einer neuen Version innerhalb des Archivssystems synchronisiert beziehungsweise aktualisiert. Bei einem Zugriff auf das Objekt wird also nicht direkt das Archivsystem angesprochen, sondern zuerst geprüft, ob das Objekt bereits in der gewünschten Version lokal vorliegt. Eine kurze Kommunikation mit dem Archivsystem findet lediglich statt, um den Status und die Konsistenz des lokal vorliegenden Objektes zu validieren.

Sichere Übertragungswege

Um die Datensicherheit und den Datenschutz zu gewährleisten, sind

sichere Übertragungswege zwischen dem Langzeitarchivierungssystem und dem zugreifenden System unerlässlich. Zwar kann eine etwaige Manipulation der Daten und Objekte durch die bereits angesprochene Checksummenüberprüfung erkannt werden, jedoch schützt dies nicht vor dem unerlaubten Zugriff Dritter auf die Objekte des Archivsystems. Dies kann sowohl über sogenanntes Abhören der Datenleitung geschehen, als auch dadurch, dass unbefugte Dritte an Zugangsdaten und Netzwerkadressen des Archivsystems gelangen. Hier ist es daher sinnvoll, mit eindeutigen Befugnissen, sicheren Übertragungsprotokollen (wie HTTPS oder SFTP) und idealerweise Signaturschlüsseln und restriktiven IP-Freigaben zu arbeiten.

Datenübernahme in ein neues Archivsystem

Ein digitales Langzeitarchivsystem sollte die Möglichkeit bieten, alle Objekte zum Zwecke einer Migration auf ein neues oder anderes Archivsystem als Gesamtpaket oder als einzelne Objekte abzurufen. Verbunden mit dem einzelnen Objekt oder dem Gesamtpaket sollten auch alle gesammelten Metadaten sein. Sie sollten nach Möglichkeit komplett in das neue Archivsystem übernommen werden.

Diese Punkte sollten bei der Planung und Umsetzung von Zugriffsstrategien auf ein Archivsystem beachtet und mit einbezogen werden. Für individuelle Lösungen werden sicherlich auch noch weitere Faktoren eine Rolle spielen. Die jeweiligen Implementierungen sind natürlich auch stark von dem verwendeten Archivsystem abhängig.

13.2 Persistent Identifier (PI) - ein Überblick

Kathrin Schroeder

Warum Persistent Identifier?

Wer eine Printpublikation bestellt, kennt i.d.R. die ISBN - eine weltweit als eindeutig angesehene Nummer. Damit kann die Bestellung sicher ausgeführt werden. Eine ähnliche Nummerierung bieten Persistent Identifier für elektronische Publikationen, die im Internet veröffentlicht werden. Damit können sehr unterschiedliche digitale Objekte wie z.B. PDF-Dokumente, Bilder, Tonaufnahmen oder Animationen dauerhaft identifiziert und aufgefunden werden.

Als "ISBN für digitale Objekte" sind die gängigen Internetadressen, die Uniform Resource Locators (URL) nicht geeignet, da diese sich zu häufig ändern.[4]

Stabile, weltweit eindeutige Identifier sind für ein digitales Langzeitarchiv unumgänglich, wie dies z.B. auch aus dem OAIS-Referenzmodell hervorgeht.

Ein von außen sichtbarer stabiler Identifier ist für die zuverlässige Referenzierung sowie für die sichere Verknüpfung von Metadaten mit dem Objekt wichtig.

Kriterien

Kriterien an PI-Systeme können sehr unterschiedlich sein. Exemplarisch sind Kriterien, die in Der Deutschen Nationalbibliothek für die Entscheidung für ein PI-System zugrunde gelegt wurden, aufgeführt.

Standardisierung

- Verankerung in internationalen Standards

4 Weiterführende Informationen zu "Adressierung im Internet und Leistungsgrenzen standortgebundener Verweise" vgl. http://www.persistent-identifier.de/?link=202

Funktionale Anforderungen

- Standortunabhängigkeit des Identifiers
- Persistenz
- weltweite Eindeutigkeit
- Der Identifier ist adressierbar und anklickbar (Resolving).
- Es kann von 1 PI gleichzeitig auf mehrere Kopien des Dokumentes (1: n-Beziehung) verwiesen werden.

Flexibilität, Skalierbarkeit

- Das PI-System ist skalierbar und
- flexibel in der PI-Anwendung selbst, d.h. es können neue Funktionalitäten hinzukommen, ohne die Konformität zum Standard zu gefährden.

Technologieunabhängigkeit und Kompatibilität

- Das PI-System ist generisch sowie protokoll- und technologieunabhängig als auch
- kompatibel mit existierenden Anwendungen und Diensten wie z.B. OpenURL, SFX, Z39.50, SRU/SRW.

Anwendung, Referenzen

- Wie verbreitet und international akzeptiert ist das PI-System?

Businessmodell und nachhaltiger Bestand

- Folgekosten (Businessmodell), Nachhaltigkeit des technischen Systems

PI-Beispiele

Nachfolgend werden die gegenwärtig als Persistent Identifier bekannten und publizierten Systeme, Spezifikationen und Standards tabellarisch vorgestellt. Zu Beginn wird das einzelne PI-System optisch hervorgehoben („Kürzel – vollständiger Name“). Die PI-Systeme sind alphabetisch geordnet.

Jede Tabelle beinhaltet die nachfolgenden Elemente:

Kurzbezeichnung	allgemein verwendete oder bekannte Abkürzung des PI-Systems
Erläuterung	kurze, allgemeine inhaltliche Erläuterungen über das Ziel sowie die Funktionalitäten des PI-Systems
Syntax	Darstellung der allgemeinen Syntax des PIs Zusätzlich wird der jeweilige PI als URN dargestellt.
Beispiel	1 oder mehrere Beispiele für einen PI
Identifizierung / Registry	kurze Angaben, was mit dem PI identifiziert wird und ob ein Registry gepflegt wird
Resolving	Wird ein Resolving unterstützt, d.h. kann der Identifier in einer klickbaren Form dem Nutzer angeboten werden
Anwender	Anwendergruppen, Institutionen, Organisationen, die das PI-System unterstützen, z.T. erfolgt dies in Auswahl
Tool-Adaption	Vorhandene Tools, Adaption in Digital Library Tools oder anderen Content Provider Systemen
Referenz	Internetquellen, Die Angabe erfolgt in Form von URLs

ARK - Archival Resource Key

Kurzbezeichnung	ARK
Erläuterung	ARK (Archival Resource Key) ist ein Identifizierungsschema für den dauerhaften Zugriff auf digitale Objekte. Der Identifier kann unterschiedlich verwendet werden: Als Link · von einem Objekt zur zuständigen Institution, · von einem Objekt zu Metadaten und · zu einem Objekt oder dessen adäquater Kopie.
Syntax	[http://NMAH/]ark:/NAAN/Name[Qualifier] NMAH: Name Mapping Authority Hostport ark: ARK-Label NAAN: Name Assigning Authority Number Name: NAA-assigned Qualifier: NMA-supported
Beispiel	http://foobar.zaf.org/ark:/12025/654xz321/s3/f8.05v.tiff Als URN: urn:ark:/12025/654xz321/s3/f8.05v.tiff

Identifizierung / Registry	- ARK-Vergabe für alle Objekte - zentrales Registry für Namensräume
Resolving	Ja, ein zentrales Register der ARK-Resolving-Dienste soll in einer „globalen Datenbank“ erfolgen, die gegenwärtig nicht von einer internationalen Agentur wie z.B. der IANA betreut wird.
Anwender	15 angemeldete Institutionen (Eigenauskunft) Darunter: California Digital Library, LoC, National Library of Medicine, WIPO, University Libraries Internet Archive, DCC, National Library of France
Tool-Adaption	Entwicklung der California Digital Library: Noid (Nice Opaque Identifier) Minting and Binding Tool
Referenz	http://www.cdlib.org/inside/diglib/ark/
Bemerkungen	Allerdings muss bei Kopien der spezif. Resolving-Service angegeben werden.

DOI – Digital Object Identifier

Kurzbezeichnung	DOI
Erläuterung	Anwendungen von Digital Object Identifiers (DOI) werden seit 1998 durch die International DOI Foundation (IDF) koordiniert. Dem DOI liegt ein System zur Identifizierung und dem Austausch von jeder Entität geistigen Eigentums zugrunde. Gleichzeitig werden mit dem DOI technische und organisatorische Rahmenbedingungen bereitgestellt, die eine Verwaltung digitaler Objekte sowie die Verknüpfung der Produzenten oder Informationsdienstleistern mit den Kunden erlauben. Dadurch wird die Möglichkeit geschaffen, Dienste für elektronische Ressourcen, die eingeschränkt zugänglich sind, auf Basis von DOIs zu entwickeln und zu automatisieren. Das DOI-System besteht aus den folgenden drei Komponenten: · Metadaten, · dem DOI als Persistent Identifier und · der technischen Implementation des Handle-Systems.

	Institutionen, die einen Dienst mit einem individuellen Profil aufbauen wollen, können dies in Form von Registration Agencies umsetzen. Das bekannteste Beispiel ist CrossRef, in dem die Metadaten und Speicherorte von Referenzen verwaltet und durch externe Institutionen weiterverarbeitet werden können. Die DOI-Foundation ist eine Non-Profit-Organisation, deren Kosten durch Mitgliedsbeiträge, den Verkauf von DOI-Präfixen und den vergebenen DOI-Nummern kompensiert werden. Die Struktur von DOIs wurde seit 2001 in Form eines ANSI/NISO-Standards (Z39.84) standardisiert, welche die Komponenten der Handles widerspiegelt.
Syntax	Präfix / Suffix
Beispiel	10.1045/march99-bunker Der Zahlencode „10“ bezeichnet die Strings als DOIs, die unmittelbar an den Punkt grenzende Zahlenfolge „1045“ steht für die vergebende Institution z.B. eine Registration Agency. Der alphanumerische String im Anschluss an den Schrägstrich identifiziert das Objekt z.B. einen Zeitschriftenartikel. Als URN: urn:doi:10.1045/march99-bunker
Identifizierung / Registry	- DOI-Vergabe für alle Objekte - zentrale Registrierung von Diensten, - Nutzer müssen sich bei den Serviceagenturen registrieren
Resolving	- Ja, Handle-System als technische Basis - Zentraler Resolving-Service - verschiedene, nicht kommunizierte dezentrale Dienste
Anwender	- 7 Registration Agencies (RA) Copyright Agency, CrossRef, mEDRA, Nielson BookData, OPOCE, Bowker, TIB Hannover - CrossRef-Beteiligte: 338 CrossRef-Nutzer - Bibliotheken (970, auch LoC) - Verlage (1528)
Tool-Adaption	Tools, welche die Nutzung von DOIs vereinfachen und die Funktionalität erweitern: http://www.doi.org/tools.html Digital Library Tools von ExLibris

Referenz	http://www.doi.org
Bemerkungen	- DOIs sind URN-konform. - kostenpflichtiger Service - gestaffelte Servicegebühren

ERRoL - Extensible Repository Resource Locator

Kurzbezeichnung	ERRoL
Erläuterung	Ein ERRoL ist eine URL, die sich nicht ändert und kann Metadaten, Content oder andere Ressourcen eines OAI-Repositories identifizieren.
Syntax	„http://errol.oclc.org/“ + <oai-identifier>
Beispiel	http://errol.oclc.org/oai:xmlregistry.oclc.org:demo/ISBN/0521555132.ListERRoLs http://errol.oclc.org/oai:xmlregistry.oclc.org:demo/ISBN/0521555132.html http://errol.oclc.org/ep.eur.nl/hdl:1765/9
Identifizierung / Registry	OAI Registry at UIUC (Grainger Engineering Library Information Center at University of Illinois at Urbana-Champaign) http://gita.grainger.uiuc.edu/registry/ListRepoIds.asp?self=1
Resolving	http-Redirect
Anwender	Nicht zu ermitteln
Tool-Adaption	DSpace
Referenz	http://errol.oclc.org/ http://www.oclc.org/research/projects/oairesolver/
Bemerkungen	Erscheint experimentell. Kein echter Persistent Identifier, da URLs aktualisiert werden müssen.

GRI – Grid Resource Identifier

Kurzbezeichnung	GRI
Erläuterung	Die Spezifikationen definieren GRI für eindeutige, dauerhafte Identifier für verteilte Ressourcen sowie deren Metadaten.
Syntax	s. URN-Syntax
Beispiel	urn:dais:dataset:b4136aa4-2d11-42bd-aa61-8e8aa5223211 urn:instruments:telescope:nasa:hubble urn:physics:colliders:cern urn:lsid:pdb.org:1AFT:1
Identifizierung / Registry	s. URN

Resolving	Im Rahmen von applikationsabhängigen Diensten wie z.B. Web-Services.
Anwender	School of Computing Science, University of Newcastle upon Tyne, Arjuna Technologies http://www.neresc.ac.uk/projects/gaf/
Tool-Adaption	http://www.neresc.ac.uk/projects/CoreGRID/
Referenz	http://www.neresc.ac.uk/ws-gaf/grid-resource/
Bemerkungen	GRI sind URN-konform.

GRid - Global Release Identifier

Kurzbezeichnung	GRid
Erläuterung	GRid ist ein System, um Releases of Tonaufnahmen für die elektronische Distribution eindeutig zu identifizieren. Das System kann Identifizierungssysteme in der Musikindustrie integrieren. Dazu gehören ein Minimalset an Metadaten, um Rechte (DRM) eindeutig zuordnen zu können.
Syntax	A Release Identifier consists of 18 characters, and is alpha-numeric, using the Arabic numerals 0 to 9 and letters of the Roman alphabet (with the exception of I and O). It is divided into its five elements in the following order: · Identifier Scheme · Issuer Code · IP Bundle Number · Check Digit
Beispiel	A1-2425G-ABC1234002-M A1 - Identifier Scheme (i.e. Release Identifier for the recording industry) 2425G - Issuer Code – (for example ABC Records) ABC1234002 - IP Bundle Number (for example an electronic release composed of a sound and music video recording, screensaver, biography and another associated video asset) M - Check Digit
Identifizierung / Registry	RITCO, an associated company of IFPI Secretariat, has been appointed as the Registration Agency.
Resolving	Resource Discovery Service
Anwender	Unklar
Tool-Adaption	unklar
Referenz	ISO 7064: 1983, Data Processing – Check Character Systems ISO 646: 1991, Information Technology – ISO 7-bit Coded Character Set for Information Exchange.

Bemerkungen	Kostenpflichtige Registrierung (150 GBP) für einen Issuer Code für 1 Jahr.

GUUID / UUID

Kurzbezeichnung	GUUID / UUID
Erläuterung	GUIDs (Globally Unique IDentifier) sind unter der Bezeichnung „UUID“ als URN-Namespace bereits bei der IANA registriert. Aufgrund des Bekanntheitsgrades werden diese erwähnt. Ein UUID (Universal Unique Identifier) ist eine 128-bit Nummer zur eindeutigen Identifizierung von Objekten oder anderen Entities im Internet. UUIDs wurden ursprünglich in dem Apollo Computer-Netzwerk, später im Rahmen der Open Software Foundation‘s (OSF), Distributed Computing Environment (DCE) und anschließend innerhalb der Microsoft Windows Platforms verwendet.
Syntax	s. URN-Syntax
Beispiel	urn:aps:node:0fe46720-7d30-11da-a72b-0800200c9a66
Identifizierung / Registry	URN-Namespace-Registry
Resolving	Kein
Anwender	Softwareprojekte
Tool-Adaption	UUID-Generatoren: http://kruithof.xs4all.nl/uuid/uuidgen http://trac.labnotes.org/cgi-bin/trac.cgi/wiki/Ruby/UuidGenerator http://sporkmonger.com/projects/uuidtools/
Referenz	http://www.ietf.org/rfc/rfc4122.txt
Bemerkungen	In der Spezifikation wird ein Algorithmus zur Generierung von UUIDs beschrieben. Wichtig ist der Ansatz, dass weltweit eindeutige Identifiers ohne (zentrale) Registrierung generiert und in unterschiedlichen Applikationen sowie verschiedene Objekttypen verwendet werden können. Wobei deutlich gesagt wird, dass UUIDs *nicht* auflösbar sind.

Handle

Kurzbezeichnung	Handle

Erläuterung	Das Handle-System ist die technische Grundlage für DOI-Anwendungen. Es ist eine technische Entwicklung der Corporation for National Research Initiatives. Mit dem Handle-System werden Funktionen, welche die Vergabe, Administration und Auflösung von PIs in Form von Handles erlauben, bereitgestellt. Die technische Basis bildet ein Protokoll-Set mit Referenz-Implementationen wie z.B. DOI, LoC.
Syntax	<Handle> ::= <Handle Naming Authority> „/“ <Handle Local Name> Das Präfix ist ein numerischer Code, der die Institution bezeichnet. Das Suffix kann sich aus einer beliebigen Zeichenkette zusammensetzen.
Beispiel	Als URN: urn:handle:10.1045/january99-bearman
Identifizierung / Registry	Zentrales Handle-Registry für die Präfixe.
Resolving	Handle-Service
Anwender	DOI-Anwender, LoC, DSpace-Anwender
Tool-Adaption	DSpace
Referenz	http://www.handle.net
Bemerkungen	Handles sind URN-konform.

InfoURI

Kurzbezeichnung	InfoURI
Erläuterung	InfoURI ist ein Identifier für Ressourcen, die über kein Äquivalent innerhalb des URI-Raumes verfügen wie z.B. LCCN. Sie sind nur für die Identifizierung gedacht, nicht für die Auflösung. Es ist ein NISO-Standard.
Syntax	„info:“ namespace „/“ identifier [„#“ fragment] info-scheme = „info“ info-identifier = namespace „/“ identifier namespace = scheme identifier = path-segments
Beispiel	**info:lccn/n78089035** **Als URN:** **urn:info:lccn/n78089035**

Identifizierung / Registry	Zentrales Registry für Namespaces
Resolving	nein
Anwender	18 Anwender: LoC, OCLC, DOI etc.
Tool-Adaption	Entwicklung für die Adaption von OpenURL-Services
Referenz	http://info-uri.info/
Bemerkungen	Zusammenarbeit mit OpenURL.

NLA - Australische Nationalbibliothek

Kurzbezeichnung	Keine vorhanden, aber die Identifier beginnen mit NLA
Erläuterung	
Syntax	Abhängig von den einzelnen Typen elektronischen Materiales werden die Identifier nach verschiedenen Algorithmen gebildet. Beispiel Collection Identifier nla.pic, nla.ms, nla.map, nla.gen, nla.mus, nla.aus, nla.arc
Beispiel	Manuscript Material <collection id>-<collection no.>-<series no.>-<item no.>-<sequence no.>-< role code>-<generation code> nla.ms-ms8822-001-0001-001-m
Identifizierung / Registry	Objekte, die archiviert werden. Es existiert ein lokales Registry.
Resolving	Ja, für die lokalen Identifier
Anwender	ANL, Zweigstellen, Kooperationspartner
Tool-Adaption	
Referenz	http://www.nla.gov.au/initiatives/persistence.html
Bemerkungen	Dies ist eine Eigenentwicklung. Es werden keine internationalen Standards berücksichtigt.

LSID - Life Science Identifier

Kurzbezeichnung	LSID
Erläuterung	Die OMG (Object Management Group) spezifiziert LSID als Standard für ein Benennungsschema für biologische Entitäten innerhalb der "Life Science Domains" und die Notwendigkeit eines Resolving-Dienstes, der spezifiziert, wie auf die Entitäten zugegriffen werden kann.

Syntax	The LSID declaration consists of the following parts, separated by double colons: • „URN“ • „LSID“ • authority identification • namespace identification • object identification • optionally: revision identification. If revision field is omitted then the trailing colon is also omitted.
Beispiel	URN:LSID:ebi.ac.uk:SWISS-PROT.accession:P34355:3 URN:LSID:rcsb.org:PDB:1D4X:22 URN:LSID:ncbi.nlm.nih.gov:GenBank.accession:NT_001063:2
Identifizierung / Registry	s. URN
Resolving	DDDS/DNS, Web-Service
Anwender	undurchsichtig
Tool-Adaption	
Referenz	http://www.omg.org/docs/dtc/04-05-01.pdf · „OMG Life Sciences Identifiers Specification.“ - Main reference page. · Interoperable Informatics Infrastructure Consortium (I3C) · Life Sciences Identifiers. An OMG Final Adopted Specification which has been approved by the OMG board and technical plenaries. Document Reference: dtc/04-05-01. 40 pages. · LSID Resolution Protocol Project. Info from IBM. · „Identity and Interoperability in Bioinformatics.“ By Tim Clark (I3C Editorial Board Member). In *Briefings in Bioinformatics* (March 2003). „Build an LSID authority on Linux.“ By Stefan Atev (IBM)
Bemerkungen	

POI - PURL-Based Object Identifier

Kurzbezeichnung	POI
Erläuterung	POI ist eine einfache Spezifikation als Resource-Identifier auf Grundlage des PURL-Systems und ist als „oai-identifier“ für das OAI-PMH entwickelt worden.

	POIs dienen als Identifier für Ressourcen, die in den Metadaten von OAI-konformen Repositories beschrieben sind. POIs können auch explizit für Ressourcen verwendet werden.
Syntax	"http://purl.org/poi/"namespace-identifier „/" local-identifier namespace-identifier = domainname-word „." domainname domainname = domainname-word [„."domainname] domainname-word = alpha *(alphanum | „-") local-identifier = 1*uric
Beispiel	http://purl.org/poi/arXiv.org/hep-th/9901001
Identifizierung / Registry	kein
Resolving	Ja, wenn dieser über das OAI-Repository bereitgestellt wird, wobei der PURL-Resolver empfohlen wird.
Anwender	unklar
Tool-Adaption	POI-Lookup-Tools http://www.rdn.ac.uk/poi/
Referenz	POI Resolver Guidelines http://www.ukoln.ac.uk/distributed-systems/poi/resolver-guidelines/ „The PURL-based Object Identifier (POI)." By Andy Powell (UKOLN, University of Bath), Jeff Young (OCLC), and Thom Hickey (OCLC). 2003/05/03. http://www.ukoln.ac.uk/distributed-systems/poi/
Bemerkungen	

PURL – Persistent URL

Kurzbezeichnung	PURL
Erläuterung	PURL (Persistent URL) wurde vom „Online Computer Library Center" (OCLC) 1995 im Rahmen des „Internet Cataloging Projects", das durch das U.S. Department of Education finanziert wurde, eingeführt, um die Adressdarstellung für die Katalogisierung von Internetressourcen zu verbessern. PURLs sind keine Persistent-Identifier, können jedoch in bestehende Standards wie URN überführt werden. Technisch betrachtet wird bei PURL der existierende Internet-Standard „HTTP-redirect" angewendet, um PURLs in die URLs aufzulösen.
Syntax	http://purl.oclc.org/OCLC/PURL/FAQ - protocol - resolver address - name

Beispiel	http://purl.oclc.org/keith/home Als URN: urn:/org/oclc/purl/keith/home
Identifizierung / Registry	Kein Registry
Resolving	ja, jedoch wird nur ein lolaker Resolver installiert.
Anwender	Keine Auskunft möglich (lt. Stuart Weibel) · OCLC · United States Government Printing Office (GPO) · LoC
Tool-Adaption	PURL-Software
Referenz	http://purl.org
Bemerkungen	· kein zentrales Registry · Die genaue Anzahl von vergebenen PURLs ist unbekannt. · Ein Test der DOI-Foundation ergab, dass nur 57% der getesteten PURLs auflösbar waren. · Experimentell von OCLC eingeführt. · Es ist keine Weiterentwicklung vorgesehen.

URN – Uniform Resource Name

Kurzbezeichnung	URN
Erläuterung	Der Uniform Resource Name (URN) existiert seit 1992 und ist ein Standard zur Adressierung von Objekten, für die eine institutionelle Verpflichtung zur persistenten, standortunabhängigen Identifizierung der Ressourcen besteht. URNs wurden mit dem Ziel konzipiert, die Kosten für die Bereitstellung von Gateways sowie die Nutzung von URNs so gering wie möglich zu halten - vergleichbar mit existierenden Namensräumen wie z.B. URLs. Aus diesem Grund wurde in Standards festgelegt, wie bereits existierende oder angewendete Namensräume bzw. Nummernsysteme einfach in das URN-Schema sowie die gängigen Protokolle wie z.B. HTTP (Hypertext Transfer Protocol) oder Schemas wie z.B. URLs integriert werden können. Der URN als Standard wird von der Internet Engineering Task Force (IETF) kontrolliert, die organisatorisch in die Internet Assigned Numbering Authority (IANA) eingegliedert ist. Sie ist für die Erarbeitung und Veröffentlichung der entsprechenden Standards in Form von „Request for Comments“ (RFCs) zuständig.

	Diese umfassen die folgenden Bereiche: · URN-Syntax (RFC 2141), · funktionale Anforderungen an URNs (RFC 1737), · Registrierung von URN-Namensräumen (z.B. RFCs 3406, 2288, 3187, NBN: 3188), · URN-Auflösungsverfahren (RFCs 3401, 3402, 3403, 3404).
Syntax	URN:NID:NISS URNs bestehen aus mehreren hierarchisch aufgebauten Teilbereichen. Dazu zählen der Namensraum (Namespace, NID), der sich aus mehreren untergeordneten Unternamensräumen (Subnamespaces, SNID) zusammensetzen kann, sowie der Namensraumbezeichner (Namespace Specific String, NISS).
Beispiel	urn:nbn:de:bsz:93-opus-59 Als URL / URI: http://nbn-resolving.de/urn:nbn:de:bsz:93-opus-59 Als OpenURL: http://[openURL-service]?identifier= urn:nbn:de:bsz:93-opus-59 Als InfoURI: info:urn/urn:nbn:de:bsz:93-opus-59 Als ARK: http://[NMAH]ark:/NAAM/urn:nbn:de:bsz:93-opus-59 Als DOI: 10.1111/urn:nbn:de:bsz:93-opus-59
Identifizierung / Registry	Überblick über den Status registrierter URN-Namensräume (unvollständig) http://www.uri.net/urn-nid-status.html
Resolving	Es gibt mehrere Möglichkeiten: - http-Redirect (Umleitung der URN zur URL) - DNS (Domain Name System)
Anwender	CLEI Code IETF IPTC ISAN ISBN ISSN

	NewsML OASIS OMA Resources XML.org Web3D MACE MPEG Universal Content Identifier TV-Anytime Forum Federated Content Government (NZ) Empfehlung: OAI 2.0: oai-identifier als URNs verwenden NBN: Finnland, Niederlande, Norwegen, Österreich, Portugal, Slovenien, Schweden, Schweiz, Tschechien, Ungarn, UK
Tool-Adaption	OPUS, DigiTool (ExLibris), Miless
Referenzen	Internetstandards: http://www.ietf.org/rfc/rfc1737.txt http://www.ietf.org/rfc/rfc2141.txt http://www.ietf.org/rfc/rfc3406.txt http://www.ietf.org/rfc/rfc288.txt http://www.ietf.org/rfc/rfc3187.txt http://www.ietf.org/rfc/rfc3188.txt http://www.ietf.org/rfc/rfc3401.txt http://www.ietf.org/rfc/rfc3402.txt http://www.ietf.org/rfc/rfc3403.txt http://www.ietf.org/rfc/rfc3404.txt URN-Prüfziffer Der Deutschen Bibliothek: http://www.pruefziffernberechnung.de/U/URN.shtml

Bemerkungen	Innerhalb der URNs sind sowohl die Integration bereits bestehender Nummernsysteme (z.B. ISBN) als auch institutionsgebundene Nummernsysteme auf regionaler oder internationaler Ebene als Namensräume möglich. Dazu zählt auch die „National Bibliography Number“ (NBN, RFC 3188), ein international verwalteter Namensraum der Nationalbibliotheken, an dem Die Deutsche Bibliothek beteiligt ist.

XRI - Extensible Resource Identifier

Kurzbezeichnung	XRI
Erläuterung	XRI wurde vom TC OASIS entwickelt. XRI erweitert die generische URI-Syntax, um „extensible, location-, application-, and transport-independent identification that provides addressability not just of resources, but also of their attributes and versions.“ zu gewährleisten. Segmente oder Ressourcen können persistent identifiziert und/oder zu adressiert werden. Die Persistenz des Identifiers wird mit den Zielen der URNs gleichgestellt.
Syntax	xri: authority / path ? query # fragment
Beispiel	xri://@example.org*agency*department/docs/govdoc.pdf XRI mit URN: xri://@example.bookstore/!(urn:ISBN:0-395-36341-1)
Identifizierung / Registry	nein
Resolving	OpenXRI.org server
Anwender	12 Förderer http://www.openxri.org/participation
Tool-Adaption	
Referenz	http://www.openxri.org/ „OASIS Releases Extensible Resource Identifier (XRI) Specification for Review.“ News story 2005-04-07. XRI Generic Syntax and Resolution Specification 1.0. Approved Committee Draft. PDF source posted by Drummond Reed (Cordance), Tuesday, 20 January 2004, 03:00pm. XRI Requirements and Glossary Version 1.0. 12-June-2003. 28 pages. [source .DOC, cache] OASIS Extensible Resource Identifier TC web site XRI TC Charter

	„OASIS TC Promotes Extensible Resource Identifier (XRI) Specification.“ News story 2004-01-19. See also „OASIS Members Form XRI Data Interchange (XDI) Technical Committee.“
Bemerkungen	

Referenzen

Beschreibung	Referenz
Überblicksdarstellung von PI-Systemen des EPICUR-Projektes	http://www.persistent-identifier.de/?link=204
PADI – Preserving Access to Digital Information	http://www.nla.gov.au/padi/topics/36.html
Nestor-Informationsdatenbank, Themenschwerpunkt: Persistente Identifikatoren	http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?show=21
ERPANET Workshop „Persistent Identifier“, 2004	http://www.erpanet.org/events/2004/cork/index.php

13.2.1 Der Uniform Resource Name (URN)

Christa Schöning-Walter

Damit digitale Objekte zitierfähig sind, müssen inhaltlich stabile Referenzen vorhanden sein, die über alle technischen und organisatorischen Veränderungen hinweg eindeutig und zuverlässig identifiziert und adressiert werden können. Grundlegende Voraussetzung ist die dauerhafte Verfügbarkeit der digitalen Objekte an sich. Deshalb ist für die Langzeitverfügbarkeit digitaler Objekte immer auch deren Speicherung in vertrauenswürdigen Archiven von zentraler Bedeutung. Persistent Identifier (PIs) haben in diesem Zusammenhang die Funktion, die eindeutige Identifizierung der ihnen zugeordneten Objekte weltweit und auf Dauer verlässlich zu gewährleisten.

Sammlung und Langzeitarchivierung von Netzpublikationen in der Deutschen Nationalbibliothek (DNB)

Mit Inkrafttreten des Gesetzes über die Deutsche Nationalbibliothek vom 22. Juni 2006 hat die DNB[5] den Auftrag der Sammlung, Erschließung, Verzeichnung und Archivierung von Netzpublikationen erhalten. Als Netzpublikationen gelten alle Darstellungen in Schrift, Bild und Ton, die in öffentlichen Netzen zugänglich gemacht werden. Dazu gehören elektronische Zeitschriften, E-Books, Hochschulprüfungsarbeiten, Forschungsberichte, Kongressschriften und Lehrmaterialien genauso wie Digitalisate alter Drucke, Musikdateien oder Webseiten. Die Grundlagen für den Aufbau eines kooperativ nutzbaren Langzeitarchivs zur Speicherung der digitalen Objekte wurden in einem vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekt (KOPAL)[6] entwickelt.

Die Langzeitarchivierung von Netzpublikationen bietet die Gewähr, dass auch die ausschließlich online veröffentlichten Werke auf Dauer verfügbar bleiben. Die Bewahrung der digitalen Objekte und die langfristige Sicherung des Zugangs stellen allerdings sehr hohe Anforderungen an die Organisation. Der Erhalt der digitalen Daten an sich muss genauso sichergestellt werden, wie die Identifizierbarkeit und Interpretierbarkeit der Inhalte. Der ständige technische Fortschritt führt zu einer sich laufend ändernden Software und Hardware und zwingt i. d. R. von Zeit zu Zeit dazu, Anpassungen des Datenträgers vorzu-

5 http://www.d-nb.de/

6 http://kopal.langzeitarchivierung.de/

nehmen. Eventuell muss sogar eine Konversion ganzer Datenbestände in eine andere Systemumgebung durchgeführt werden, damit die Benutzbarkeit der Inhalte erhalten bleibt.
Im Lebenszyklus digitaler Objekte kann sich neben dem Ort der Speicherung also immer wieder auch das technische Format verändern. Die DNB bewahrt alle Formate auf, die im Zuge ihrer Maßnahmen zur Langzeitarchivierung entstehen. Die Informationen über die notwendigen Systemvoraussetzungen für die Benutzung (Hardware und Software, Dateiformat, u. a.) werden dabei als Metadaten zusammen mit dem digitalen Objekt gespeichert.
In diesem Zusammenhang ist es notwendig, dass alle Netzpublikationen, die archiviert werden sollen, einen PI besitzen. Der PI ist ein permanenter Name, der einer Netzpublikation über ihren gesamten Lebenszyklus hinweg zugeordnet bleibt. Er hat die Funktion, ein digitales Objekt (und die dazu gehörenden Metadaten) unabhängig vom Speicherort über die Systemgrenzen und Systemwechsel hinweg in allen seinen Repräsentationen auf Dauer eindeutig zu identifizieren.
Die DNB verwendet dafür den Uniform Resource Name (URN). Digitale Objekte, denen bisher noch kein URN zugeordnet wurde, erhalten spätestens bei der Erschließung in der DNB einen eindeutigen Namen, der selbstverständlich auch in anderen Zusammenhängen (z. B. in anderen Archiven) zur Identifizierung der so gekennzeichneten Objekte genutzt werden kann.

Das Schema der Uniform Resource Names (URNs)

Die funktionale Spezifikation von URNs gehört zu den Basiskonzepten, die Anfang der 1990er Jahre im Zusammenhang mit dem Entwurf der Architektur für das World Wide Web (WWW) entwickelt wurden. URNs sind eine bestimmte Form der Uniform Resource Identifier (URIs). URIs identifizieren die Ressourcen im Internet. Das URN-Schema beschreibt den Rahmen für die Identifizierung von Informationsressourcen mittels weltweit gültiger eindeutiger Bezeichnungen (Namen).
Einschlägige Entwicklungen des Internets werden durch die Internet Assigned Numbers Authority (IANA)[7] kontrolliert. Die Arbeitsgruppen der Internet Engineering Task Force (IETF)[8], eine Organisation, die IANA zugeordnet ist, treiben die Weiterentwicklung voran und legen die de facto-Standards fest. Beschreibungen und Empfehlungen sind in der Form so genannter Requests for Comments (RFCs) veröffentlicht.

7 http://www.iana.org/

8 http://www.ietf.org/

Mit dem Dokument RFC 1737[9] (Functional Requirements for URNs, 1994) wurden schon sehr früh die grundlegenden Anforderungen an das URN-Schema spezifiziert. RFC 2141[10] (URN Syntax, 1997) beschreibt etwa 2 Jahre später u. a. die Ziele, die mit der Entwicklung dieses PIs verfolgt wurden:

> *Uniform Resource Names (URNs) are intended to serve as persistent, location-independent resource identifiers and are designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space. Therefore, the URN syntax provides a means to encode character data in a form that can be sent in existing protocols, transcribed on most keyboards, etc.*

Das URN-Schema ist also ganz bewusst sehr offen konzipiert worden, um bereits vorhandene Bezeichnungssysteme oder Standardnummern (beispielsweise ISBNs), andere Schemata (beispielsweise URLs) oder gängige Protokolle (beispielsweise http) direkt integrieren zu können. Man wollte einerseits Unabhängigkeit vom Ort der Speicherung eines Objekts und dem verwendeten Zugriffsprotokoll erreichen, andererseits aber auch den Aufwand für die Bereitstellung von Gateways so gering wie möglich halten.
Die Einbettung neuer oder auch schon vorhandener Namensschemata in das URN-Schema erfolgt durch die Registrierung von Namensräumen bei IANA.[11] Ein Namensraum kennzeichnet gewissermaßen den Geltungsbereich eines URNs und definiert die Menge der Objekte, welche mittels der angewendeten Systematik identifiziert und adressiert werden sollen. So können – bei Einhaltung des durch das generische Schema definierten Rahmens – durchaus sehr spezifische Konventionen festgelegt werden. Bei der Registrierung sollte allerdings auch immer die Verlässlichkeit des PIs nachgewiesen werden (RFC 3406, URN Namespace Definition Mechanisms)[12].
IANA verzeichnet gegenwärtig (Stand: Februar 2008) 36 verschiedene Namensräume. Dazu gehören u. a.:

- issn – International Serials Number (RFC 3044),
- isbn - International Standards Books Number (RFC 3187),
- isan – International Standard Audiovisual Number (RFC 4246),
- nbn – National Bibliography Number (RFC 3188),
- pin – Personal Internet Name für Personen und Organisationen (RFC 3043),
- uuid – Universally Unique Identifiers für verteilte Softwaresysteme (RFC 4122).

9 http://www.ietf.org/rfc/rfc1737.txt

10 http://www.ietf.org/rfc/rfc2141.txt

11 http//www.iana.org/assignments/urn-namespaces

12 http://www.ietf.org/rfc/rfc3406.txt

Zur Auflösung von URNs in Zugriffsadressen werden Resolvingdienste zwischengeschaltet (RFC 2276, URN Resolution) [13]. Die Resolver verwalten Metadaten zu allen im System registrierten Objekten. Um die Objekte zu lokalisieren, werden URNs i. d. R. über ein Register in Uniform Resource Locator (URLs) umgewandelt.
Die zu einem URN-Namensraum gehörenden Resolvingdienste müssen in der Lage sein, registrierte Informationsressourcen solange nachzuweisen, wie Exemplare des jeweiligen Objekts oder Referenzen auf das Objekt irgendwo existieren. Entsprechend ist die Persistenz eines URNs auch immer unmittelbar davon abhängig, ob eine stabile und leistungsfähige Infrastruktur vorhanden ist, welche die zugehörigen Dienste zur Vergabe, Verwaltung und Auflösung registrierter Objekte zuverlässig und langfristig erbringen kann.
Zusammenfassend kann gesagt werden, dass jedes URN-Schema die folgenden Anforderungen erfüllen muss:

- Gültigkeit des Namens weltweit,
- Eindeutigkeit des Namens weltweit,
- Persistenz: Benutzbarkeit des Namens auf Dauer,
- Skalierbarkeit: das Schema muss beliebig viele Namen aufnehmen können,
- Ausbau-/Erweiterungsfähigkeit: die Systematik muss eine Weiterentwicklung oder Migration zulassen,
- Übertragbarkeit: andere regelkonforme Bezeichnungssysteme müssen eingebettet werden können,
- Unabhängigkeit: die beteiligten Institutionen selbst legen die Namenskonventionen fest,
- Auflösbarkeit: die Verfügbarkeit von Resolvingdiensten muss auf Dauer gewährleistet sein.

Die National Bibliography Number (NBN)

Zu den bei IANA registrierten Namensräumen zählt auch die NBN. Sie wurde entwickelt, um die rasant anwachsende Zahl und Vielfalt digitaler Publikationen – beispielsweise elektronische Zeitschriften, Hochschulschriften, Forschungsberichte, Lehr- und Lernmaterialien, u. a. – in den Nationalbibliografien besser verzeichnen zu können. Das Konzept beruht auf einer Initiative der Conference of Directors of National Libraries (CDNL) und der Conference of European National Librarians (CENL). Es wurde von Juha Hakala (Finnische

13 http://www.ietf.org/rfc/rfc2276.txt

Nationalbibliothek) beschrieben (RFC 3188, 2001) [14].
Die NBN ist international gültig. Wie in Deutschland übernehmen i. Allg. auch in anderen Ländern die Nationalbibliotheken das Management des Namensraums auf nationaler Ebene. In das internationale Netzwerk der aktiv beteiligten und untereinander vernetzten Partner sind die meisten skandinavischen Länder, einige baltische Staaten, die Schweiz, Österreich und Italien eingebunden.
Die DNB betreibt einen Resolving-Dienst für Deutschland, Österreich und die Schweiz.[15] Zu diesem Dienst gehört auch ein Internetportal, das Informationen und Werkzeuge für die Benutzer zur Verfügung stellt. Der Aufbau erfolgte im Rahmen eines vom BMBF geförderten Modellprojekts (EPICUR) [16]. Die Konventionen und Qualitätskriterien des Dienstes sind in der URN-Strategie der DNB dokumentiert.
Mit diesem Namensraum steht für Autoren, Verlage, Bibliotheken, Archive, Forschungseinrichtungen und andere Institutionen ein kooperativ anwendbares Verfahren zur Registrierung und Auflösung von PIs für ihre elektronischen Publikationen zur Verfügung. Die Nachteile einer standortbezogenen Identifizierung lassen sich damit überwinden. Eine Verweisung auf die genaue Speicheradresse eines Objekts im Internet ist i. a. nicht auf Dauer benutzbar. Demgegenüber behalten URN-basierte Referenzen in Publikationen, Bibliothekskatalogen, Bibliografien oder Portalen auch dann ihre Gültigkeit, wenn sich der Ort der Speicherung verändert (beispielsweise bei technischen Umstrukturierungen oder bei der Verlagerung eines digitalen Archivs). Der zwischengeschaltete Resolver ermöglicht es, den Aufwand zur Pflege ungültig gewordener Speicheradressen relativ gering zu halten, weil lediglich der Eintrag im Register korrigiert werden muss.
Die Persistenz des Identifiers ist allerdings keine Eigenschaft an sich. Sie kann nur in enger Kooperation aller am System beteiligten Institutionen gewährleistet werden und erfordert

- die Vergabe und Registrierung eindeutiger Namen für die Informationsressourcen,
- eine leistungsfähige Infrastruktur zur Auflösung der Namen (Resolving),
- die Einhaltung der festgelegten Regeln,
- unterstützende organisatorische und technische Maßnahmen zur Qualitätssicherung,
- und die dauerhafte Verfügbarkeit der digitalen Objekte an sich.

14 http://www.ietf.org/rfc/rfc3188.txt

15 http://nbn-resolving.de/

16 http://www.persistent-identifier.de/

Die URN-Struktur

URNs sind streng hierarchisch strukturiert und gliedern sich in einen Präfix und einen Suffix.
RFC 2141 beschreibt die allgemeine Syntax eines URNs:

urn:[NID]:[SNID]-[NISS]

Präfix:

- NID Kennzeichnung des Namensraums (Namespace Identifier)
- SNID optional können zusätzlich Unternamensräume definiert werden (Subnamespace Identifier)

Suffix:

- NISS Kennzeichnung des Objekts (Namespace Specific String)

Das Präfix identifiziert den Geltungsbereich (Namensraum) des URNs sowie die für die Verlässlichkeit und Auflösung des einzelnen Namens verantwortlichen Institutionen.
Ein URN, der mit urn:nbn:de beginnt, drückt immer aus, dass es sich um eine NBN handelt, die in Deutschland vergeben wurde und die über den Resolver der DNB aufgelöst werden kann.
Die auf internationaler Ebene eingeleitete hierarchische Strukturierung kann auf nationaler Ebene durch Gliederung in Unternamensräume weiter fortgesetzt werden. Institutionen oder Personen, die URNs vergeben wollen, können einen Unternamensraum beantragen. Die Registrierung von Unternamensräumen erfolgt in Deutschland bei der DNB. Bibliotheken wählen i. d. R. ein Kennzeichen, das sich aus dem Namen des Bibliotheksverbundes und dem Bibliothekssigel zusammensetzt. Für Institutionen oder Personen, die sich nicht in die organisatorische Struktur der Bibliotheksverbünde einordnen (wie zum Beispiel Verlage, Forschungseinrichtungen, Verbände oder Firmen), wird i. d. R. eine vierstellige Zahlenkombination als Identifikator festgelegt.
Das Suffix eines URNs schließlich ist eine Zeichenfolge zur eindeutigen Identifizierung der Informationsressource selbst und kann aus Buchstaben, Zahlen und Sonderzeichen bestehen.
Die in Deutschland vergebenen URNs im Namensraum nbn:de haben den folgenden Aufbau:

urn:nbn:de:[Unternamensraum]-[eindeutige Identifikation des Objekts][Prüfziffer]

Beispiel 1:

Metadaten-Kernset im Format ONIX[17], hrsg. von der DNB

urn:nbn:de:101-2007072707

Präfix:	
urn:nbn:de	Kennzeichen des Auflösungsdienstes
101	Kennzeichen der URN-Vergabestelle; hier: DNB
Suffix:	
200707270	Zeichenfolge zur eindeutigen Identifikation des Objekts; hier: Aufnahmedatum
7	Prüfziffer (wird automatisch generiert)

Beispiel 2:

Hans-Werner Hilse, Jochen Kothe: Implementing Persistent Identifiers (hrsg. vom Consortium of European Research Libraries)

urn:nbn:de:gbv:7-isbn-90-6984-508-3-8

Präfix:	
urn:nbn:de	Kennzeichen des Auflösungsdienstes
gbv:7	Kennzeichen der URN-Vergabestelle; hier: SUB Göttingen
Suffix:	
isbn-90-6984-508-3-	Zeichenfolge zur eindeutigen Identifikation des Objekts hier: ISBN
8	Prüfziffer (wird automatisch generiert)

Auch innerhalb des Namensraums nbn:de können also lokal oder global bereits eingeführte Namensschemata wie z. B. die ISBN als Identifikatoren für ein Objekt verwendet werden.

17 Online Information Exchange, Datenformat zum Austausch von bibliografischen und Produktdaten im Buchhandel

Die Auflösung von URNs

URNs werden in nationalen und internationalen Nachweissystemen (z. B. Bibliografien, Kataloge und Suchmaschinen) nachgewiesen und sind über bibliografische Austauschformate transportierbar.

Nach Möglichkeit sollten URNs bereits im Zuge der Publikation vergeben werden, weil sie dann direkt in die Publikation mit eingebettet und so veröffentlicht werden können. Durch die hierarchische Struktur bleibt die Eindeutigkeit der Namen auch bei einer stark dezentral organisierten Anwendung des URN-Schemas gewährleistet.
Damit URNs auflösbar sind, müssen sie zuvor im Resolver registriert worden sein. Erst danach ist ein URN für die Identifizierung und Adressierung einer Informationsressource benutzbar.
Ein URN verweist auf mindestens einen URL. In der Regel werden mehrere Kopien und unterschiedliche Präsentationsformate (zum Beispiel HTML, PDF, JPEG) eines Objekts verwaltet. Typischerweise verweist der Resolver sowohl auf die Repräsentationen des Objekts vor Ort – z. B. auf den Dokumentenserver der Hochschule, des Verlags oder der Forschungseinrichtung – als auch auf eine Kopie in einem Langzeitarchiv (z. B. auf das Langzeitarchiv der DNB).

Für die zeitnahe Übermittlung von Namen und Standortadressen digitaler Objekte an den URN-Resolver der DNB stehen Frontendsysteme, Transferschnittstellen, standardisierte Datenaustauschformate und automatisierte Übertragungsverfahren (Harvesting) zur Verfügung.
Bei Verwaltung mehrerer URLs zu einem URN existiert ein Standardverhalten des Resolvers. Vorrangig wird der URL mit der höchsten Priorität aufgelöst. Das kann z. B. der Volltext einer Publikation in einem bestimmten Format (beispielsweise PDF) sein oder eine Webseite mit einer Beschreibung des Objektes (Frontdoor). Falls dieser URL vorübergehend oder dauerhaft nicht erreichbar ist, wird der URL mit der nächsten Priorität benutzt. Die Reihenfolge wird bei der Registrierung des URNs festgelegt. Die Auflösbarkeit eines URNs auf Dauer kann allerdings nur dann gewährleistet werden, wenn auch mindestens eine Kopie in einem vertrauenswürdigen Langzeitarchiv vorhanden ist. Ansonsten kann ein URN eventuell auch ungültig werden. Der Name bleibt dennoch erhalten und dem dann nicht mehr vorhandenen Objekt zugeordnet.
Ein URN dient ausschließlich zur Identifizierung eines einzelnen Objekts. Der Resolver kann keine Informationen über den Kontext verarbeiten, zum Beispiel Informationen über die Struktur einer elektronischen Zeitschrift mit mehreren

Beispiel 3: Suche im Katalog der DNB

Beispiel 4: Anzeige der zu einem URN registrierten Adressen

Bänden und darin enthaltenen einzelnen Artikeln. Allerdings darf das Objekt, auf das sich ein URN bezieht, mehrere inhaltlich selbständige Beiträge beinhalten. So umfasst beispielsweise ein URN, der sich auf die Titelseite einer Zeitschrift bezieht, alle veröffentlichten Bände der Zeitschrift. Gleichzeitig können aber auch alle adressierbaren Teilobjekte ihrerseits einen URN besitzen (z.B. die einzelnen Bände einer Zeitschrift oder sogar die einzelnen Artikel).
Um einen URN aufzulösen, muss der zugehörige Resolvingdienst gefunden werden. URNs können – mit der Adresse des Resolvers zu einer http-Adresse verknüpft – in den Browser eingegeben werden. Der dahinterliegende Resolvingdienst führt in diesem Fall die Standardauflösung durch und realisiert den direkten Zugriff auf das digitale Objekt. Die Angabe nur des URNs genügt i. d. R. nur dann, wenn spezielle Plugins installiert sind.

Beispiel 5: Auflösung eines URNs über die http-Adresse

http://nbn-resolving.de/urn:nbn:de:bvb:703-opus-3845

⇩

http://opus.ub.uni-bayreuth.de/volltexte/2008/384

Benutzer, die einen URN auflösen wollen, können dafür aber auch die Webseite des Resolvers benutzen.[18]

Beispiel 6: Auflösung eines URNs über die Webseite des Resolvers in der DNB

URNs sind ein Teil der Internet-Architektur

Alle Uniform Resource Identifier (URIs), die im Internet bzw. im WWW verwendet werden – so also auch der URN – müssen dem aktuellen Standard für URIs, RFC 3986[19] (URI: Generic Syntax, 2005), entsprechen. Die Basisarchitektur des WWW mit URIs als Grundkonzept für die Identifizierung jeglicher Ressourcen (RFC 1630, Universal Resource Identifiers in WWW)[20] stammt bereits aus dem Jahre 1994 und wurde von Tim Berners Lee entworfen. Das

18 http://www.persistent-identifier.de/?link=610

19 http://www.ietf.org/rfc/rfc3986.txt

20 http://www.ietf.org/rfc/rfc1630.txt

Prinzip gilt in gleicher Art und Weise für physikalische wie auch für abstrakte Ressourcen (Zugriff auf Dateien oder Webseiten, Aufruf von Webservices, Zustellung von Nachrichten, u. a.).
Der jetzt vorliegende Standard spezifiziert den grundsätzlichen Aufbau eines URIs. Die einzelnen Schemata können allerdings weiterhin sehr unterschiedlich sein. Gekennzeichnet wird jedes Schema durch seinen Namen, gefolgt von einem Doppelpunkt.
IANA verzeichnet gegenwärtig (Stand: Februar 2008) mehr als 60 verschiedene permanente URI-Schemata[21]. Neben dem URN gehören dazu u. a.:

- ftp – File Transfer Protocol
- http – Hypertext Transfer Protocol
- info – InfoURI
- mailto – E-mail-Adresse
- z39.50r – Z39.50 Retrieval
- z39.50s - Z39.50 Session

Die nachfolgende Tabelle gibt abschließend einen zusammenfassenden Überblick über die wichtigsten IETF-Empfehlungen[22], die in ihrer Gesamtheit den URN als einen Uniform Resource Identifier beschreiben:

Request for Comments (RFC)	**Thema**	**Status**	**Datum**
Grundlage: das URI-Schema			
RFC 1630	Universal Resource Identifiers in WWW	Informational	1994
RFC 3986	Uniform Resource Identifier Generic Syntax	Standards Track	2005
URNs: Funktionale Anforderungen			
RFC 1737	Functional Requirements for Uniform Resource Names	Informational	1994

21 http://www.iana.org/assignments/uri-schemes
22 http://www.ietf.org/rfc.html

URN-Syntax

RFC 2141	URN Syntax	Standards Track	1997

Definition von Namensräumen

RFC 2288	Using Existing Bibliographic Identifiers as Uniform Resource Names	Informational	1998
RFC 3187	Using ISBNs as URNs	Informational	2001
RFC 3188	Using National Bibliography Numbers (NBNs) as URNs	Informational	2001
RFC 3406	URN Namespace Definition Mechanisms	Best Current Practise	2002

Auflösungsverfahren (Resolving)

RFC 2169	A Trivial Convention for using HTTP in URN Resolution	Experimental	1997
RFC 2276	Architectural Principles of Uniform Resource Name Resolution	Informational	1998
RFC 2483	URI Resolution Services	Experimental	1999
RFC 3401 RFC 3402 RFC 3403 RFC 3404	Dynamic Delegation Discovery System	Standards Track	2002

Literatur

Hans-Werner Hilse, Jochen Kothe: Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. Consortium of European Research Libraries. European Commission on Preservation and Access, 2006. urn:nbn:de:gbv:7-isbn-90-6984-508-3-8

EPICUR: Uniform Resource Name (URN) – Strategie der Deutschen Nationalbibliothek (2006). urn:nbn:de:1111-200606299

Kathrin Schroeder: EPICUR. In: Dialog mit Bibliotheken, 17 (2005) 1, S. 58 – 61

Kathrin Schroeder: Persistent Identifiers im Kontext der Langzeitarchivierung. In: Dialog mit Bibliotheken, 16 (2004) 2, S. 11 – 14

13.2.2 Der Digital Objekt Identifier (DOI) und die Verwendung zum Primärdaten-Management

Dr. Jan Brase

Der Digital Object Identifier (DOI)

Der Digital Object Identifier (DOI) wurde 1997 eingeführt, um Einheiten geistigen Eigentums in einer interoperativen digitalen Umgebung eindeutig zu identifizieren, zu beschreiben und zu verwalten. Verwaltet wird das DOI-System durch die 1998 gegründete International DOI Foundation (IDF)[23].
Der DOI-Name ist ein dauerhafter persistenter Identifier, der zur Zitierung und Verlinkung von elektronischen Ressourcen (Texte, aber Primärdaten oder andere Inhalte) verwendet wird. Über den DOI-Namen sind einer Ressource aktuelle und strukturierte Metadaten zugeordnet.
Ein DOI-Name unterscheidet sich von anderen, gewöhnlich im Internet verwendeten Verweissystemen wie der URL, weil er dauerhaft mit der Ressource als Entität verknüpft ist und nicht lediglich mit dem Ort, an dem die Ressource platziert ist.
Der DOI-Name identifiziert eine Entität direkt und unmittelbar, also nicht eine Eigenschaft des Objekts (eine Adresse ist lediglich eine Eigenschaft des Objekts, die verändert werden und dann ggf. nicht mehr zur Identifikation des Objekts herangezogen werden kann).
Das IDF-System besteht aus der „International DOI Foundation“ selbst, der eine Reihe von Registrierungsagenturen („Registration Agencies “; RA) zugeordnet sind . Für die Aufgaben einer RA können sich beliebige kommerzielle oder nicht kommerzielle Organisationen bewerben, die ein definiertes Interesse einer Gemeinschaft vorweisen können, digitale Objekte zu referenzieren.

Technik

Das DOI-System baut technisch auf dem Handle-System auf. Das Handle System wurde seit 1994 von der US-amerikanischen Corporation for National Research Initiatives (CNRI[24]) als verteiltes System für den Informationsaustausch entwickelt. Handles setzen direkt auf das IP-Protokoll auf und sind eingebettet in ein vollständiges technisches Verwaltungsprotokoll mit festgelegter Prüfung der Authentizität der Benutzer und ihrer Autorisierung. Durch das Handle-Sys-

23 http://www.doi.org/

24 http://www.cnri.reston.va.us/ bzw. http://www.handle.net

tem wird ein Protokoll zur Datenpflege und zur Abfrage der mit dem Handle verknüpften Informationen definiert. Diese Informationen können beliebige Metadaten sein, der Regelfall ist aber, dass die URL des Objektes abgefragt wird, zu dem das Handle registriert wurde. Weiterhin stellt CNRI auch kostenlos Software zur Verfügung, die dieses definierte Protokoll auf einem Server implementiert (und der damit zum sog. Handle-Server wird).
Ein DOI-Name besteht genau wie ein Handle immer aus einem Präfix und einem Suffix, wobei beide durch einen Schrägstrich getrennt sind und das Präfix eines DOI-Namens immer mit „10." Beginnt. Beispiele für DOI-Namen sind:

doi:10.1038/35057062
doi:10.1594/WDCC/CCSRNIES_SRES_B2

Die Auflösung eines DOI-Namens erfolgt nun über einen der oben erwähnten Handle-Server. Dabei sind in jedem Handle-Server weltweit sämtliche DOI-Namen auflösbar. Dieser große Vorteile gegenüber anderen PI-Systemen ergibt sich einerseits durch die eindeutige Zuordnung eines DOI-Präfix an den Handle-Server, mit dem dieser DOI-Name registriert wird und andererseits durch die Existenz eines zentralen Servers bei der CNRI, der zu jedem DOI-Präfix die IP des passenden Handle-Servers registriert hat. Erhält nun ein Handle-Server irgendwo im Netz den Auftrag einen DOI-Namen aufzulösen, fragt er den zentralen Server bei der CNRI nach der IP-Adresse des Handle-Servers, der den DOI-Namen registriert hat und erhält von diesem die geforderte URL.

DOI-Modell

Die Vergabe von DOI-Namen erfolgt wie oben erwähnt nur durch die DOI-Registrierungsagenturen, die eine Lizenz von der IDF erwerben. Dadurch wird sichergestellt, dass jeder registrierte DOI-Namen sich an die von der IDF vorgegebenen Standards hält. Diese Standards sind als Committee Draft der ISO Working Group TC46 SC9 WG7 (Project 26324 Digital Object Identifier system) veröffentlicht und sollen ein anerkannter ISO Standard werden. Zum Stand 12/07 gibt es 8 DOI-Registrierungsagenturen, die teilweise kommerzielle, teilweise nicht-kommerzielle Ziele verfolgen. Bei den Agenturen handelt es sich um

- Copyright Agency Ltd[25], CrossRef[26], mEDRA[27], Nielsen BookData[28] und

25 http://www.copyright.com.au/
26 http://www.crossref.org/
27 http://www.medra.org/
28 http://www.nielsenbookdata.co.uk/

R.R. Bowker[29] als Vertreter des Verlagswesens,
- Wanfang Data Co., Ltd[30] als Agentur für den Chinesischen Markt,
- OPOCE (Office des publications EU)[31], dem Verlag der EU, der alle offiziellen Dokumente der EU registriert
- Technische Informationsbibliothek (TIB) als nicht-kommerzielle Agentur für Primärdaten und wissenschaftliche Information

Dieses Lizenz-Modell wird häufig gleichgesetzt mit einer kommerziellen Ausrichtung des DOI-Systems, doch steht es jeder Registrierungsagentur frei, in welcher Höhe sie Geld für die Vergabe von DOI-Namen verlangt. Auch muss berücksichtigt werden, dass – anders als bei allen anderen PI-Systemen – nach der Vergabe von DOI-Namen durch die Verwendung des Handle-Systems für das Resolving- bzw. für die Registrierungs-Infrastruktur keine weiteren Kosten entstehen,

Die TIB als DOI Registrierungsagentur für Primärdaten

Der Zugang zu wissenschaftlichen Primärdaten ist eine grundlegende Voraussetzung für die Forschungsarbeit vor allem in den Naturwissenschaften. Deshalb ist es notwendig, bestehende und zum Teil auch neu aufkommende Einschränkungen bei der Datenverfügbarkeit zu vermindern.
Traditionell sind Primärdaten eingebettet in einen singulären Forschungsprozess, ausgeführt von einer definierten Gruppe von Forschern, geprägt von einer linearen Wertschöpfungskette:

Experiment	⇒	Primärdaten	⇒	Sekundärdaten	⇒	Publikation
		Akkumulation		Datenanalyse		Peer-Review

Durch die Möglichkeiten der neuen Technologien und des Internets können einzelne Bestandteile des Forschungszyklus in separate Aktivitäten aufgeteilt werden (Daten-Sammlung, Daten-Auswertung, Daten-Speicherung, usw.) die von verschiedenen Einrichtungen oder Forschungsgruppen durchgeführt werden können. Die Einführung eines begleitenden Archivs und die Referenzierung einzelner Wissenschaftlicher Inhalte durch persistente Identifier wie einen DOI-Namen schafft die Möglichkeit anstelle eines linearen Forschungsansatzes, den Wissenschaftlerarbeitsplatz einzubinden in einen idealen Zyklus der Information und des Wissens (siehe Abbildung 13.2.2.1), in dem durch Zentrale Datenarchive als Datenmanager Mehrwerte geschaffen werden können und so für alle Datennutzer, aber auch für die Datenautoren selber ein neuer Zugang

29 http://www.bowker.com/

30 http://www.wanfangdata.com/

31 http://www.publications.eu.int/

zu Wissen gestaltet wird.

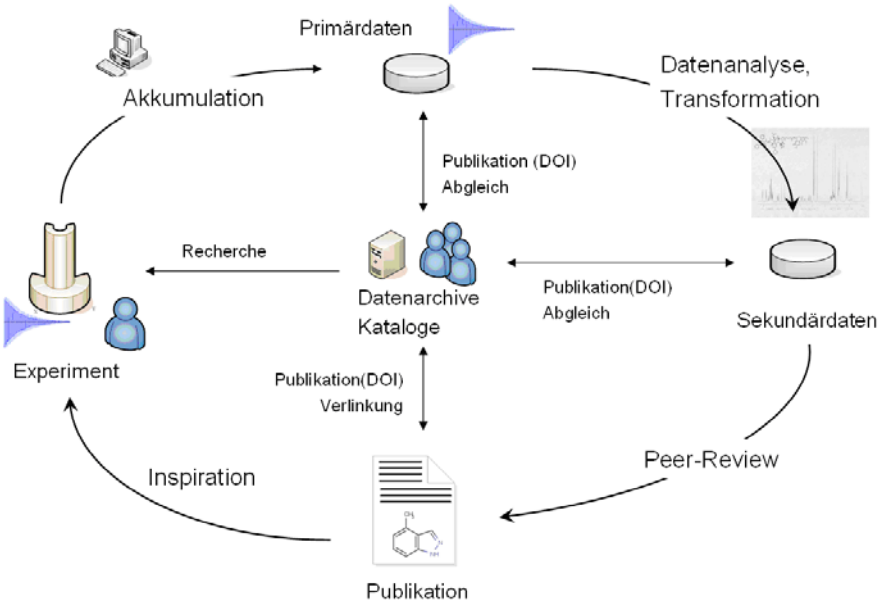


Abbildung 13.2.2.1: Ein idealer Zyklus der Information und des Wissens

Der DFG-Ausschuss „*Wissenschaftliche Literaturversorgungs- und Informationssysteme*“ hat 2004 ein Projekt[32] gestartet, um den Zugang zu wissenschaftlichen Primärdaten zu verbessern. Aus diesem Projekt heraus ist die TIB seit Mai 2005 weltweit erste DOI-Registrierungsagentur für wissenschaftliche Daten. Beispielhaft im Bereich der Geowissenschaften werden Primärdatensätze registriert. Die Datensätze selber verbleiben bei den lokalen Datenzentren und die TIB vergibt für jeden Datensatz einen DOI-Namen.
Der Datensatz wird somit eine eigene zitierfähige Einheit. Mittlerweile wurden über dieses System über 500.000 Datensätze mit einer DOI versehen und zitierfähig gemacht. Die Metadatenbeschreibungen der Datensätze werden zentral an der TIB gespeichert. Diese Beschreibungen enthalten alle Angaben, die nach ISO 690-2 (ISO 1997) zur Zitierung elektronischer Medien verlangt werden.

32 http://www.std-doi.de

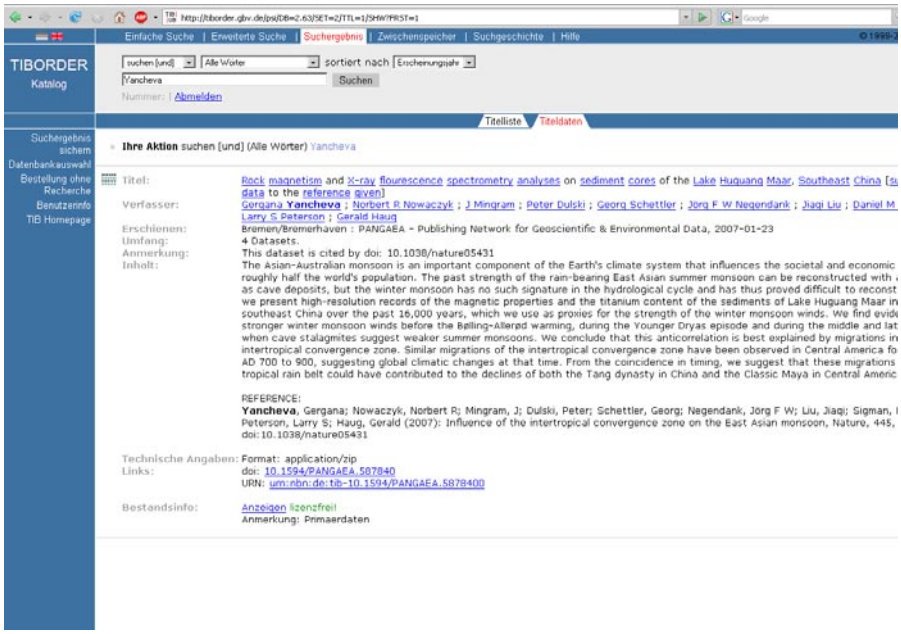


Abbildung 13.2.2.2: Anzeige eines Primärdatensatzes im Online-Katalog der TIB Hannover

Zusätzlich werden Sammlungen oder Auswertungen von Primärdatensätzen auch in den Katalog der TIB aufgenommen. Die Anzeige eines Primärdatensatzes im Katalog der TIB sehen sie in Abbildung 13.2.2..2.

Die DOI Registrierung erfolgt bei der TIB immer in Kooperation mit lokalen Datenspeichern als sog. Publikationsagenten, also jenen Einrichtungen die weiterhin für Qualitätssicherung und die Pflege und Speicherung der Inhalte, sowie die Metadatenerzeugung zuständig sind. Die Datensätze selber verbleiben bei diesen lokalen Datenzentren, die TIB speichert die Metadaten und macht alle registrierten Inhalte über eine Datenbank suchbar. (Brase, 2004; Lautenschlager et al., 2005)

Für die Registrierung von Datensätzen wurde an der TIB ein Webservice eingerichtet. Komplementär wurden bei den Publikationsagenten entsprechende Klienten eingerichtet, die sowohl eine automatisierte als auch manuelle Registrierung ermöglichen. In allen Datenzentren sind die SOAP[33]-Klienten vollständig in die Archivierungsumgebung integriert, so dass zusätzlicher Arbeitsaufwand für die Registrierung entfällt. Mithilfe dieser Infrastruktur sind bisher problemlos mehrere hunderttausend DOI Namen registriert worden. Das System baut

33 SOAP steht für *Simple Object Access Protocol*, ein Netzwerkprotokoll, mit dessen Hilfe Daten zwischen Systemen ausgetauscht werden können

seitens der TIB auf dem XML-basierten Publishing-Framework COCOON von Apache auf. Dazu wurde COCOON um eine integrierte Webservice-Schnittstelle erweitert, wodurch die Anbindung von weiterer Software überflüssig wird. Die modulare Struktur des Systems erlaubt es, dieses auf einfache Weise auf alle weiteren Inhalte, die mit DOI Namen registriert werden, anzupassen.

Status

Die DOI-Registrierung von Primärdaten ermöglicht eine elegante Verlinkung zwischen einem Wissenschaftlichen Artikel und den im Artikel analysierten Primärdaten. Artikel und Datensatz sind durch die DOI in gleicher Weise eigenständig zitierbar.

So wird beispielsweise der Datensatz:
G.Yancheva, . R Nowaczyk et al (2007)
Rock magnetism and X-ray flourescence spectrometry analyses on sediment cores of the Lake Huguang Maar, Southeast China, PANGAEA
doi:10.1594/PANGAEA.587840

in folgendem Artikel zitiert.
G. Yancheva, N. R. Nowaczyk et al (2007)
Influence of the intertropical convergence zone on the East Asian monsoon
Nature 445, 74-77
doi:10.1038/nature05431

Mittlerweile hat die TIB ihr Angebot auch auf andere Inhaltsformen ausgeweitet.[34] Als Beispiele seien hier genannt:

- doi:10.1594/EURORAD/CASE.1113 in Kooperation mit dem European Congress for Radiology (ECR) wurden über 6.500 medizinische Fallstudien registriert.
- doi:10.2312/EGPGV/EGPGV06/027-034 in Kooperation mit der European Association for Computer Graphics (Eurographics) wurden über 300 Artikel (Graue Literatur) registriert.
- doi:10.1594/ecrystals.chem.soton.ac.uk/145 Gemeinsam mit dem Projekt eBank des UK Office for Library Networking wurden erstmals DOI

34 Weitere Informationen zu den Aufgaben der TIB als DOI-Registrierungsagentur und dem Nachweis von Primärdaten durch DOI-Namen sind auf den Internetseiten der TIB zu finden http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/ und http://www.tib-hannover.de/de/spezialsammlungen/primaerdaten/

Namen für Kristallstrukturen vergeben.
- doi:10.2314/CERN-THESIS-2007-001 in Kooperation mit dem CERN werden DOI Namen für Berichte und Dissertationen vergeben
- doi:10.2314/511535090 Seit Sommer 2007 vergibt die TIB auch DOI Namen für BMBF Forschungsberichte.

DOI-Namen und Langzeitarchivierung

Die Referenzierung von Ressourcen mit persistenten Identifiern ist ein wichtiger Bestandteil jedes Langzeitarchivierungskonzeptes. Der Identifier selber kann natürlich keine dauerhafte Verfügbarkeit sicherstellen, sondern stellt nur eine Technik dar, die in ein Gesamtkonzept eingebunden werden muss. Ein Vorteil der DOI ist hier sicherlich einerseits der zentrale Ansatz durch die überwachende Einrichtung der IDF, der die Einhaltung von Standards garantiert und andererseits die breite Verwendung der DOI im Verlagswesen, das an einer dauerhaften Verfügbarkeit naturgemäß interessiert ist. In sehr großen Zeiträumen gerechnet gibt es natürlich weder für die dauerhafte Existenz der IDF noch der CNRI eine Garantie. Allerdings ist die Technik des Handle Systems so ausgelegt, dass eine Registrierungsagentur jederzeit komplett selbstständig die Auflösbarkeit ihrer DOI-Namen sicherstellen kann.

Literatur

Brase, J., 2004. Using Digital Library Techniques - Registration of Scientific Primary Data. Lecture Notes in Computer Science, 3232: 488-494.

International Organisation for Standardisation (ISO). ISO 690-2:1997 Information and documentation, TC 46/SC 9

Lautenschlager, M., Diepenbroek, M., Grobe, H., Klump, J. and Paliouras, E., 2005. World Data Center Cluster „Earth System Research" - An Approach for a Common Data Infrastructure in Geosciences. EOS, Transactions, American Geophysical Union, 86(52, Fall Meeting Suppl.): Abstract IN43C-02.

Uhlir, Paul F. ,2003 The Role of Scientific and Technical Data and Information in the Public Domain, National Academic Press, Washington DC

14 Technischer Workflow

Reinhard Altenhöner

14.1 Einführende Bemerkungen und Begriffsklärungen

Immer dann, wenn Termini und Methoden zur Beschreibung und zur Modellierung von Abläufen aus einem anderen Umfeld in den Kontext eines spezifischen Themas oder spezialisierter Abläufe eingeführt werden, entsteht Bedarf für einen der eigentlichen Beschäftigung mit dem Gegenstand vorgehende Definitions- und Klärungsschritt. Konkret: Die Langzeitarchivierung als relativ neuem Arbeitsgebiet, in dem bislang der Schwerpunkt stark auf forschungsnahen oder gar experimentellen Ansätzen lag, wird beim Übergang zu produktiven Systemen und operativen Ablaufproblemen mit neuen Aufgabenstellungen konfrontiert: Jetzt geht es um umfassende Arbeitsabläufe, um die massenhafte Prozessierung von (automatisierten) Arbeitsschritten und es scheint sinnvoll, hier auf das Erfahrungswissen und die Methodik aus anderen Arbeitsbereichen und Geschäftsfeldern zurückzugreifen.

Und da der Bewusstheitsgrad, mit dem Arbeitsprozesse im kommerziellen Kontext – oft über aufwändige Beratungsdienste durch einschlägige Anbieter

- organisatorisch und technisch modelliert werden, sehr hoch ist, lohnt es sich, zunächst auf das methodische und begriffliche Umfeld einzugehen, aus dem heraus die Terminologie rund um „Workflow" entstanden ist. Das gilt sicher generell für das Thema (technische) Prozessorganisation, um so mehr aber für das Arbeitsfeld der Langzeitarchierung, das insbesondere in Bibliotheken, Archiven und Museen zunehmend bedeutender wird, das aber bislang bis auf wenige Ausnahmen noch nicht in größerem Umfang etabliert und in die allgemeinen Arbeitsabläufe generell integriert ist. Es folgen daher hier zunächst einige einführende Begriffsklärungen, die dann im nächsten Schritt dann für die konkrete Thematik Langzeitarchivierung methodisch/konzeptionell aufgegriffen werden, um schließlich in einem weiteren Schritt den bislang erreichten Praxisstand an einigen Beispielen etwas eingehender zu betrachten. Ergänzend noch der Hinweis, dass in diesem Handbuch zwischen dem organisatorischen und dem technischen Workflow differenziert wird.

Der Begriff des Workflow wird im Deutschen im Allgemeinen mit dem Begriff des Geschäftsprozesses gleichgesetzt. Aus der abstrahierende Beschreibung von Einzelfällen entsteht die Basis dafür, Abläufe systematisch als Arbeits- oder Geschäftsprozess zu beschreiben und zum Beispiel daraus Schulungsmaterial zu generieren, aber auch Schwachstellen zu identifizieren oder neue Fallgruppen zu integrieren. Mit der darunter liegenden Ebene der Arbeitsschritte – der Arbeitsprozess (work process) ist als eine geordnete Folge von Arbeitsschritten definiert - wird bereits ein relativ hoher Detaillierungsgrad erreicht, der es erlaubt, Abläufe differenziert zu verstehen.

Mit Hilfe einer regelbasierten Beschreibung der Abläufe ergibt sich aber auch die Möglichkeit, Geschäftsprozesse zu planen, bewusst in systematischer Weise einzugreifen, Teile oder ganze Abläufe neu zu modellieren, also die Abläufe zu steuern, zu „managen". In diesen Prozessen werden Dokumente, Informationen oder auch Aufgaben von einem Teilnehmer zum anderen gereicht, die dann nach prozeduralen Regeln bearbeitet werden. In klassischer Definition wird der Workflow übrigens mit der teilweisen oder vollständigen Automatisierung eines Geschäftsprozesses gleich gesetzt.[1]

Enger auf den Bereich der öffentlichen Verwaltung bezogen und so auch in Bibliotheken gebraucht ist der Begriff des Geschäftsgangs, hier häufig festgemacht am Bearbeitungsobjekt, in der Regel Büchern oder auch Akten und dem Weg dieser Objekte durch die einzelnen Phasen seiner Bearbeitung. Gemeint ist

1 Martin (1999), S. 2.

hier – trotz der verwaltungstypischen Fokussierung auf die bearbeiteten Objekte – der Arbeitsablauf/Geschäftsprozess als Gesamtheit aller Tätigkeiten zur Erzeugung eines Produktes bzw. zur Erstellung einer Dienstleistung.[2]

Das Workflow-System bezeichnet dagegen die IT-gestützte integrierte Vorgangsbearbeitung, in der Datenbank, Dokumentenmanagement und Prozessorganisation in einem Gesamtkonzept abgebildet werden.[3] Diese Abläufe werden also technisch unterstützt, wenn nicht sogar überhaupt mit Hilfe technischer Werkzeuge und Methoden betrieben. Aber auch die Modellierung von Geschäftsprozessen selbst kann toolunterstützt erfolgen, solche Geschäftsprozeßmanagement-Tools dienen der Modellierung, Analyse, Simulation und Optimierung von Prozessen. Die entsprechenden Applikationen unterstützen in der Regel eine oder mehrere Methodiken, ihr Funktionsspektrum reicht von der Ist-Aufnahme bis zur Weitergabe der Daten an ein Workflow-Management-System. Im Mittelpunkt stehen dabei Organisation, Aufgaben bzw. Ablauf der Aufgaben und die zugrundeliegenden Datenmodelle. Mit der Schnittstelle solcher Tools zum Beispiel zu Workflow-Management-Systemen beschäftigt sich die Workflow-Management-Coalition[4], die sich insbesondere die Austauschbarkeit der Daten und damit die Interoperabilität zwischen unterschiedlichen, zum Teil spezialisierten Tools durch entsprechende Standardisierungsanstrengungen auf die Fahnen geschrieben hat.

Der Begriff des technischen Workflows schließlich wird im Allgemeinen primär für die Abläufe verwandt, die einen hohen Automatisierungsgrad bereits haben oder wenigstens das Potential dazu. Entsprechend bezeichnet man mit dem Begriff des Technischen Workflow-Management die Systeme, die durch eine geringe Involviertheit von Menschen und eine hohe Wiederholbarkeit bei geringen Fehlerquoten gekennzeichnet sind.

Damit ist klar, dass der Begriff des technischen Workflow im Kontext der Langzeitarchivierung geradezu programmatischen Charakter hat, da angesichts der großen Objektmengen und ihrer prinzipiell gegebenen Eigenschaften als digitale Publikation, ein hoher Automatisierungsgrad besonders bedeutsam ist.

---

2 Verwaltungslexikon (2008), Eintrag Workflow. Damit der englischen Ausgangsbedeutung des Begriffs folgend.

3 Verwaltungslexikon (2008), aaO.

4 http://www.wfmc.org/

14.2 Workflow in der Langzeitarchivierung: Methode und Herangehensweise

Die allmähliche Einführung der Langzeitarchivierung in das reguläre Auftragsportfolio von Bibliotheken und anderen Kulturerbeeinrichtungen mit immer höheren Bindungsquoten von Personal und anderen Ressourcen erzeugt(e) zunächst neue, häufig isolierte und händisch durchgeführte Abläufe, verändert aber auch in einer ganzheitlichen Betrachtung Arbeitsabläufe und die sie modellierenden Geschäftsprozesse. So ist schon für sich die Einspielung von Daten in ein Langzeitarchiv ein komplexer Vorgang, in dem eine ganze Reihe von auf einander bezogenen bzw. von einander abhängenden Aktivitäten ablaufen. Vor allem aber die zunehmende Relevanz der technischen und operativen Bewältigung der Aufgabe verlangt nach einer systematischen Modellierung der Geschäftsprozesse, also dem Einstieg in ein systematisches Workflowmanagement. Es gilt allerdings festzustellen, dass selbst in Einrichtungen, die bereits seit einigen Jahren Erfahrungen mit dem Betrieb von Langzeitarchiven und ihrer Integration in die jeweilige Systemlandschaft gesammelt haben, häufig noch isolierte Bearbeitungsketten ablaufen, die zudem keinesfalls wirklichen Vollständigkeitsgrad haben, also alle Anforderungs- /arbeitsfelder abdecken und außerdem vielfach noch manuelle Eingriffe erfordern, insbesondere auf dem Gebiet des Fehlermanagements.

Diese Feststellung bedeutet aber auch, dass der Erfahrungshorizont zum technischen Workflow insgesamt noch relativ gering ist, also hier noch konkrete Erfahrungen vor allem im Umgang mit großen Mengen und insbesondere auch im automatisierten Qualitätsmanagement gewonnen werden müssen. Insofern hat die Beschäftigung mit dem technischen Workflow derzeit noch viele theoretische Elemente und hat propädeutischen Charakter.

Vor allem in einer Situation, in der verschiedene (bereits existente und neu entwickelte) Arbeitsprozesse ineinander greifen und auch verschiedene Organisationseinheiten an ein und demselben Vorgang beteiligt sind, ist die Modellbildung ein Beitrag zur umfassenden Optimierung. Damit befinden sich Bibliotheken, Archive und Museen in einer Situation, die man mit den Anstrengungen der Privatwirtschaft Anfang der 1990er Jahre vergleichen kann, als dort die Modellierung von Geschäftsprozessen unter verschärften Wettbewerbs- und Kostendruckbedingungen systematischer als zuvor angegangen wurde. Auch wenn im öffentlich finanzierten Umfeld in besonderem Maße historisch geprägte Orga-

nisationsformen gegeben sind, die eine vorgangsbezogene Sicht erschweren, führt an der grundsätzlichen Anforderung der Neu-Modellierung aus systematischer Sicht kein Weg vorbei. Diese wird im Umfeld des technischen Workflow immer stark auch von der informationstechnischen Entwicklungsseite getrieben sein, denn Ziel der Geschäftsprozessmodellierung ist ihre technische Abbildung.

Übergeordnete Ziele dieses Herangehens, also der systematischen Modellierung und eines methodenbewussten Workflowmanagements sind:

- Verbesserung der Prozessqualität
- Vereinheitlichung der Prozesse
- schnellere und zuverlässigere Bearbeitung von Aufträgen (extern und intern)
- Reduzierung der Durchlaufzeiten
- Kostenreduktion
- Verbesserte Verfügbarkeit von Information / Dokumentation
- Erhöhte Prozessflexibilität
- Erhöhung der Transparenz der Prozesse (Statusermittlung, Dokumentation von Entscheidungen), Qualitätssicherung
- Automatische Eingriffsmöglichkeiten: Dokumentation, Eskalation bei Zeitüberschreitungen, Verteilung von Aufgaben und Verantwortlichkeiten
- Vermeidung von Redundanz, mangelnder Aktualität und Inkonsistenz durch Mehrfachschritte

Natürlich lassen sich kleine isolierte Prozesse oder Prozesselemente durch individuelle Programmierung jeweils neu umsetzen. Dies geschah in der Vergangenheit vielfach für einzelne Objektklassen oder auch einzelne Datenübergabe- oder –tauschprozesse. Aber schon beim Zusammenführen bzw. Hintereinandersetzen der einzelnen Teilschritte bedarf es einer Gesamtlogik für das Management des Ablaufs dieser Schritte. Fehlt diese Logik, verbleiben letztlich viele immer wieder manuelle neu anzustoßende Teilkonstrukte mit dazu häufig proprietären „Konstruktions“elementen. Schon insofern ist die systematische Analyse verschiedener wiederkehrender Arbeitsabläufe ein sinnvoller Ansatz, um so zur Modellierung auch komplexer Vorgänge aus dem Bereich der Langzeitarchivierung zu kommen.
Erst auf dieser Basis wird es möglich, Services zu definieren, die wieder verwendbar sind, weil sie Arbeitsschritte abbilden, die in verschiedenen Umfeldern vorkommen, beispielsweise das Aufmachen eines Bearbeitungsfalls für ein Objekt und die IT-gestützte Verwaltung verschiedener Be-/Verarbeitungsschritte

dieses Objekts. In dieser Perspektive entsteht der Geschäftsprozess für eine Klasse von Objekten aus der Zusammenfügung verschiedener Basisservices, die miteinander interoperabel sind. Das diese Herangehensweise sehr stark mit dem Modell der Serviceorientierten Architektur (SOA) bei der Entwicklung IT-basierter Lösungen korrespondiert, ist dabei kein Zufall. Voraussetzung dafür ist aber wie angesprochen die Modellierung der Arbeits- oder Geschäftsprozesse, die vorgeben, welche Services wann und wie gebraucht werden. Die Prozessmodellierung bildet also die Basis für die Implementierung, die Prozesse selbst dienen der Orchestrierung, dem Zusammenspiel und der Aufeinandereinstimmung der Services. In einem optimalen (Infrastruktur)Umfeld können so die Arbeitsschritte als kleinere Einheit eines Geschäftsprozesses verschiedene Services lose zusammenbringen.

Die Informatik hat für die Modellierung und Notation von Geschäftsprozessen verschiedene methodische Herangehensweisen entwickelt, zum Beispiel die Ereignisgesteuerten Prozessketen (EPK), eine von Scheer und Mitarbeitern entwickelte Sprache zur Modellierung von Geschäftsprozessen[5] und vor allem die Unified Modeling Language (UML) der Object Management Group (OMG), die in der Praxis heute dominierende Modellierungssprache für die Modellierung von Daten, Verhalten, Interaktion und Aktivitäten.[6]

Zur vorbereitenden Modellierung von technischen Abläufen in der Langzeitarchivierung wird man sich zunächst am OAIS-Modell orientieren, das die prinzipiellen Aufgaben im Umfeld der Langzeitarchivierung in funktionaler Perspektive beschreibt und an anderer Stelle dieser Enzyklopädie ausführlich beschrieben wird.[7]

Einzelne Funktionen lassen sich so vor der Folie bisher bereits gemachter Erfahrungen allgemein beschreiben. Beispiele für diese übergreifenden Basisprozesse sind (ich nenne nur Beispiele für unmittelbar aus dem Kontext der Langzeitarchivierung heraus relevante Prozesse):

- Plattform- und Systemübergreifendes Taskmanagement
- Daten- und Objekttransfer-Mimik (z.B. OAI, ORE)
- Extraktion und Generierung von Metadaten (METS, LMER)
- Validierung von Dokumentformaten (z.B. JHOVE)
- Persistente Adressierung und Zugriffsmanagement auf Objektebene

5 Keller (1992)

6 OMG Infrastructure (2007) und OMG Superstructure (2007)

7 Hier Link auf den entsprechenden Artikel ???

- Speicherprozesse
- ID-Management
- Inhaltsauswahl / Basisrecherche
- Migrationsprozesse / Formatkonvertierungen
- On-the-fly-Generierung einer Bereitstellungsumgebung

14.3 Technisches Workflowmanagement in der Praxis: Erfahrungen und Ergebnisse

Insgesamt ist wie dargelegt der Umfang praktischer Erfahrungen noch begrenzt. Wichtige Erkenntnisse konnte sowohl in der technischen Workflowentwicklung als auch in der praktischen Umsetzung die niederländische Nationalbibliothek sammeln, doch auch die Deutschen Nationalbibliothek, die nach einer Gesetzesnovelle Mitte des Jahres 2006 die Zuständigkeit für die Erhaltung der Langzeitverfügbarkeit deutscher Online – oder Netzpublikationen erhalten hat, steht vor sehr konkreten Herausforderungen, die derzeit zu einer umfassenden Reorganisation des technischen Workflow führen.[8] Mit dem Inkrafttreten des neuen Gesetzes und der damit verbundenen deutlich erweiterten Verpflichtung, die Aufgabe der Langzeitarchivierung zu erfüllen, stellt sich hier die Frage in einer neuen Dimension: Wie wird die Bibliothek die neuen Abläufe organisieren, welche technischen Methoden und Anwendungen werden im Massenverfahren eingesetzt? Da gleichzeitig die alten Arbeitsabläufe und –verfahren weiterlaufen, stellt sich die Frage der Integration in ganz anderer Weise. Zwar ist die Bibliothek in der glücklichen Situation, für die neuen Aufgaben zusätzliche Ressourcen erhalten zu haben, doch würden diese nicht eine nahtlose Imitation des organisatorisch-operativen Workflows auf Basis der existierenden Systeme abdecken – das ergibt sich schon aus den Mengen, um die es geht.

Königliche Bibliothek der Niederlande (KB): Die KB betreibt seit dem Jahr 2003 das OAIS-kompatibles Archivierungssystem DIAS der Firma IBM operativ und hat im Laufe der gewonnenen Erfahrungen insbesondere organisatorisch eine ganze Reihe von Anpassungen unternommen.[9] Technisch gesehen wurde eine auch in der KB weitgehend isolierte gesonderte Entwicklung aufgesetzt, die über eine nur geringe Anbindung an die sonstigen Abläufe der Bibliothek bietet. Schwerpunkt liegt auf dem Ingest-Prozess, also dem Einspielen des in der Regel von Verlagen bereitgestellten publizierten Materials in das Archiv. Dieses erfolgt weitgehend automatisiert und es ist der Niederländischen Nationalbibliothek sehr schnell gelungen, die Fehlerquoten auf minimale Prozentbereiche zu drücken. Inzwischen sind mehr als zehn Millionen Objekte eingespielt, darunter auch einige komplexe Objekte wie historische CD-ROMs.

8 Es sei angemerkt, dass es eine ganze Reihe von weiteren Publikationen zum Thema gibt. So stellte etwa Clifton (2005) Worksflows der australischen Nationalbibliothek vor; diese beziehen sich allerdings auf die manuelle Behandlung von Objekten mittels einzelner Tools.

9 KB (2008)

Für alle Objekte – es handelt sich in der weit überwiegenden Zahl um PDF-Dateien – gilt, dass in der eigentlichen Langzeitarchivumgebung rudimentäre Metadateninformationen gespeichert sind; die bibliographischen Informationen werden über ein Recherchesystem der KB zur Verfügung gestellt.

Insgesamt ist es der KB gelungen, den technischen Workflow relativ unkompliziert und damit effizient und für hohe Durchsatzmengen geeignet zu halten. Dies war auch deswegen möglich, weil die Zahl der Lieferanten in das System in den Niederlanden klein ist, da wenige große Verlage bereits einen überwiegenden Anteil am Publikationsvolumen der Niederlande haben.

In Deutschland stellt sich die Situation anders dar: Hier bestimmen viele in einer zum Teil noch sehr traditionell geprägten Veröffentlichungslandschaft Verleger das Bild. Ausgangspunkt für die Deutsche Nationalbibliothek war eine Situation, in der für die Verarbeitung von Online-Dokumenten bereits eine Vielzahl von mehr oder weniger halbautomatische Verfahren für für Netzpublikationen, Online-Dissertationen und weitere Materialien existierte. Diese historisch gewachsenen Strukturen standen nebeneinander, d.h. – nicht untypisch für Gedächtnisorganisationen im öffentlichen Kontext – die einzelne Objektklasse war der definitorische Ausgangspunkt für einen hochspezialisierten Workflow. Ziel war und ist daher die Schaffung eines automatischen, einheitlichen Verfahrens mit der Übergabe der Archivobjekte an das im Rahmen des Projekts kopal entstandene Archivsystem und die dort entstandenen Verfahren.[10] Sowohl Ingest wie auch der Zugriff auf die Objekte sehen die Übergabe aus der Langzeitarchivlösung kopal am Arbeitsplatzrechner vor oder an das neu entstehende Bereitstellungssystem. Dabei sind zahlreiche Arbeitsbereiche in der DNB involviert: neben dem bibliographischen System sind dies die Fachbereiche, externe Ablieferer, aber auch die für die digitalen Dienste der DNB Verantwortlichen. Insofern ist hier vieles noch offen und ein Werkstattbericht mag dies illustrieren:[11]

Für den Transfer und das Angebot von Objekten auf elektronischen Materialien auf physischen Datenträgern (d.h. CD- bzw. DVD-Veröffentlichungen) existiert ein älterer, segmentierte Workflow, die nun aufgrund der Anforderungen seitens Archivsystem und künftiger Bereitstellung anzupassen sind. Nach Erstellung der Images der Daten und einer Analyse des vorhandenen Materials wurde daher ein Änderungs- und Ergänzungsvorschlag für den inte-

10 kopal (2008)

11 Wollschläger (2007), S. 18ff.

grierten Workflow dieser Materialgruppe erarbeitet.

Ebenso wird der Workflow für genuin online vorliegende Netzpublikationen unter Einbeziehung der Anforderungen der Langzeitarchivierung neu gestaltet und auf die Schnittstellen des Archivsystems angepasst. Dabei ergeben sich eine ganze Reihe von Problemen: So entsprechen fortlaufende Publikationen (vor allem elektronische Zeitschriften-Artikel) und die künftigen zu archivierenden Objekte häufig nicht der aktuellen Abbildung im Online-Katalog. Bibliografische Metadaten von Archivobjekten müssen aber künftig im bibliographischen System abgebildet werden, um einen einheitlichen Zugang zu gewährleisten. Dazu müssen eine Festlegung von Erschließungsvarianten und ein Mapping von Archivobjekten auf Katalogobjekte erfolgen, letztlich also eine klare Definition der Granularität von Objekten und ihrer Abbildung getroffen werden.

Das URN-Management in der DNB wurde bereits erweitert und vor allem technisch so weiterentwickelt, dass eine Einbindung in andere Arbeitszusammenhänge erfolgen kann. Da jedes Objekt zum Einspielen in das Archiv einen Persistent Identifier benötigt, erfolgt für bereits gesammelte Objekte ohne URN eine retrospektive Vergabe der URN. Alle neuen Objekte müssen entweder mit URN geliefert werden bzw. bei Eingang/Bearbeitung einen URN erhalten, was dem künftigen Verfahren entspricht.

Wesentliche Voraussetzungen für die Einbindung des Archivs in die Geschäftsumgebung der Institution liegen mittlerweile vor oder werden gerade geschaffen. Insbesondere die Kernelemente des Produktionssystem laufen, das produktive Einspielen von Material wurde und wird erprobt, nötige Weiterentwicklungen (z.B. noch fehlende Module zur Auswertung von Dateiformaten) wurden und werden ermittelt und Änderungen / Anpassungen in diversen Workflows der traditionellen Bearbeitung wurden bereits angestoßen. Weitere Aufgaben betreffen in hohem Maße die Übergabe des kopal-Systems eine ständige Arbeitseinheit sowie die retrospektive Aufarbeitung des früher bereits in die Bibliothek gelangten Materials.

Hinter diesen Bemühungen steht der Anspruch, die neuen, mit der Gesetzesnovelle übernommenen Aufgaben, die weit über das Arbeitsfeld der Langzeitarchivierung hinausgehen, in einem ganzheitlichen technischen Workflow abzubilden. In dessen Mittelpunkt stehen aktuell die Übernahme von elektronischen Objekten mit möglichst breiter Nachnutzung vorhandener Metainformationen

und die Integration der Abläufe in die Arbeitsumgebung der DNB.

Die praktischen Erfahrungen an der DNB insbesondere für diesen Bereich belegen den besonderen Bedarf für eine bewusste Modellierung der Geschäftsprozesse, die in der Vergangenheit häufig nur unvollkommen gelungen ist. Im Ergebnis standen isolierte, von nur wenigen Personen bediente und bedienbare Abläufe mit einem hohem manuellen Eingriffs- und Fehlerbehandlungsbedarf. Ohne dass heute bereits ein komplettes Profil der zukünftigen technischen Workflow-Umgebung zitierfähig vorliegt, kann doch gesagt werden, dass ein methodisch bewusstes, in enger Kooperation von Bedarfsträger und Informationstechnik ablaufendes Vorgehen zu deutlich klareren Vorstellungen darüber führt, wie die wesentlichen Arbeitsschritte exakt aussehen und wie sie adäquat so abgebildet werden, dass die entstehenden Services auch langfristig und damit über ihren aktuellen Entstehungshintergrund hinaus genutzt werden.

Dass dabei für eine technische Arbeitsumgebung besondere Anforderungen an die Flexibilität und die Orientierung an offenen Standards gelten, liegt auf der Hand und hat wesentlich die Entwicklungsleitlinien für kopal mitbestimmt.[12]

Literatur

Clifton, Gerard: Safe Havens In A Choppy Sea: Digital Object Management Workflows At The National Library of Australia (2005), Beitrag zur iPRES - International Conference on Preservation of Digital Objects, Göttingen (September 15, 2005). In: http://rdd.sub.uni-goettigen.de/conferences/ipres05/download/Safe%20Havens%20In%20A%20Choppy%20Sea%20Digital%20Object%20Management%20Workflows%20At%20The%20National%20Library%20of%20Australia%20-%20Gerard%20Clifton.pdf (Zugriff 6.1.2008)

Keller, Gerhard / Nüttgens, Markus / Scheer, August-Wilhelm (1992): Semantische Prozessmodellierung auf der Grundlage „Ereignisgesteuerter Prozessketten (EPK)". In: A.-W. Scheer (Hrsg.): Veröffentlichungen des Instituts für Wirtschaftsinformatik, Heft 89, Saarbrücken. Online in: http://www.iwi.uni-sb.de/Download/iwihefte/heft89.pdf (Zugriff 28.12.2007)

Königliche Bibliothek der Niederlande (KB): How the e-Depot works (2008) In: http://www.kb.nl/dnp/e-depot/dm/werking-en.html (Zugriff 6.1.2008)

Königliche Bibliothek der Niederlande (KB): The e-Depot system (DIAS) (2008) In: http://www.kb.nl/dnp/e-depot/dias-en.html (Zugriff 6.1.2008)

Kopal (2008): Projekthompage. In: http://kopal.langzeitarchivierung.de/ (Zugriff am 6.1.2008)

12 kopal (2008a)

Kopal (2008a): kopal: Ein Service für die Langzeitarchivierung digitaler Informationen. In: http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf (Zugriff am 6.1.2008)

Martin, Norbert (1999): Und wie kommt die Dissertation auf den Server? Gedanken zum Workflow. Vortrag auf der IuK-Tagung „ Dynamic Documents „ vom 22.-24.3.1999 in Jena. In: http://edoc.hu-berlin.de/epdiss/jena3/workflow.pdf (Zugriff am 4.1.2008)

OMG Infrastructure (2007). UML Infrastructure Specification, v2.1.2. OMG document formal/07-11-04. In: http://www.omg.org/docs/formal/07-11-04.pdf (Zugriff am 4.1.2008)

OMG Superstructure (2007). UML Superstructure Specification, v2.1.2. OMG document formal/07-11-02. In: http://www.omg.org/docs/formal/07-11-02.pdf (Zugriff am 4.1.2008)

Stapel, Johan: The KB e-Depot. Workflow Management in an Operational Archiving Environment (2005). Beitrag zur iPRES - International Conference on Preservation of Digital Objects, Göttingen (September 15, 2005). In: http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Workflow%20Management%20In%20An%20Operational%20Archiving%20Environment%20-%20Johan%20Stapel.pdf (Zugriff 6.1.2007)

Verwaltungslexikon (2008) - Management und Reform der öffentlichen Verwaltung (2008) In: http://www.olev.de/w.htm (Zugriff am 4.1.2008)

Wollschläger, Thomas (2007): „kopal goes live“. In: Dialog mit Bibliotheken 19 (2007), H.2, S. 17 – 22

Workflow Management Coalition (2008) – Website. In: http://www.wfmc.org/ (Zugriff am 5.1.2008)

15 Anwendungsfelder in der Praxis

Einleitung

Regine Scheffel

Die vorangegangenen Kapitel über Strategien, Modelle, Standards u. a. vermitteln den (derzeitigen) Kenntnisstand, der notwendig ist, um kompetent Probleme der Langzeitarchivierung und Langzeitverfügbarkeit anzupacken. Vielfach treten jedoch Anforderungen zutage, die Praktikerinnen und Praktiker in (Kulturerbe-)Institutionen nicht kurzfristig selbst klären, ändern oder erfüllen können (z. B. policies, Organisationsmodelle oder Hardwareumgebung). Dennoch stehen sie unter Handlungsdruck, um die digitalen Objekte in ihrem Veranwortungsbereich nutzbar zu erhalten. Hier setzt das folgende Kapitel an, das konkrete Anwendungsfelder der genannten Aspekte (z. B. Formate) in der Praxis vorstellt.

Diese Anwendungsfelder beziehen sich nicht auf Handlungsfelder in Bibliotheken, Museen, Archiven oder Forschungseinrichtungen (z. B. Publikation), sondern auf den Umgang mit den unterschiedlichen Medienarten wie Text, Bild und Multimedia in seinen diversen Ausprägungen. Darüberhinaus werden Langzeitarchivierung und Langzeitverfügbarkeit komplexer digitaler Material-

sammlungen thematisiert, die über den Medienmix hinaus weitere spezifische Anforderungen stellen, z. B. Websites, wissenschaftliche Rohdaten oder Computerspiele.

15.1 Textdokumente

Karsten Huth

Definition

Die Definition des Begriffs Textdokument im Bereich der Langzeitarchivierung bzw. die Antwort auf die Frage: "Was ist ein Textdokument?", ist nicht einfach zu beantworten. Kommen doch zwei Ebenen eines digitalen Objekts für eine Definitionsgrundlage in Frage[1]. Auf der konzeptuellen Ebene liegt ein Textdokument genau dann vor, wenn das menschliche Auge Text erkennen, lesen und interpretieren kann. Diese Anforderung kann auch eine Fotografie, bzw. das Bild eines Textes erfüllen. Auf der logischen Ebene eines digitalen Objektes, der Ebene der binären Codierung und Decodierung liegt ein Textdokument genau dann vor, wenn innerhalb des Codes auch Textzeichen codiert sind und dadurch Gegenstand von Operationen werden (z.B. Kopieren und Verschieben, Suchen nach bestimmten Worten und Wortfolgen, Ersetzen von bestimmten Zeichenfolgen usw.).

Da ein Archiv seine Archivobjekte generell auf der konzeptuellen Ebene betrachten muss, insbesondere da sich die technikabhängige logische Ebene im Laufe der Zeit durch Migration grundsätzlich ändert[2], soll für dieses Kapitel die erste Definition zur Anwendung kommen:

> *Ein Textdokument liegt genau dann vor, wenn das menschliche Auge Text erkennen, lesen und interpretieren kann.*

Diese Definition ermöglicht die Verwendung von Dateiformaten zur Speicherung von Bildinformationen ebenso wie die speziell auf Textverarbeitung ausgerichteten Formate. Welchen Formattyp ein Archiv zur Speicherung wählt, hängt von den wesentlichen Eigenschaften des Archivobjekts ab. Die wesentlichen Eigenschaften eines digitalen Archivobjekts müssen vom Archiv bei oder bereits vor der Übernahme des Objekts in das Archiv festgelegt werden und ergeben sich gemäß den Vorgaben des OAIS größtenteils aus den Ansprüchen und Möglichkeiten der Archivnutzer.[3]

---

1 vgl. Funk, Stefan, Kap 9.1 Digitale Objekte

2 vgl. Funk, Stefan, Kap 12.2 Migration

3 Consaltative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information Sytem: Blue Book*. Washington, DC. Page 3-4

Archivierung von Textdokumenten mit Bildformaten:

Die Archivierung von Textdokumenten in Bildformaten empfiehlt sich genau dann, wenn der optische Eindruck eines Textdokuments eine der wesentlichen Eigenschaften des Archivobjekts ist, welches auf das Genaueste erhalten werden muss. Solche Fälle ergeben sich z.B. bei der Digitalisierung von amtlichem Schriftgut, bei der anschließend das Original aus Papier vernichtet wird, während man die digitale Fassung archiviert. Da bei diesen Objekten das originale Schriftbild sowie von Hand aufgetragene Zeichen (z.B. Anmerkungen, Unterschriften und Paraphen) für die dauerhafte korrekte Interpretation des Archivobjektes unbedingt erhalten werden müssen, ist die Speicherung als Bild der beste Weg. In einem Bildformat sind in der Regel nur Informationen über Bildpunkte und ihre jeweiligen Farb- und Helligkeitswerte in einem Raster verzeichnet (Bitmap-Grafik). Diese Formate beinhalten von sich aus keinerlei Informationen über den abgebildeten Text. Deshalb kann man in einer solchen Datei nicht nach bestimmten Textstellen suchen, Textinhalte herauskopieren oder verschieben. Die Unveränderlichkeit der inhaltlichen und optischen Darstellung ist für ein Archiv von Vorteil.

Eine Abhandlung zu möglichen Bildformaten im Archiv befindet sich im Kapitel 15.2 „Bilddokumente“.[4] Bildformate werden in diesem Kapitel nicht weiter thematisiert.

Archivierung von Textdokumenten mit Textformaten:

Die Archivierung von Textdokumenten in Textformaten empfiehlt sich genau dann, wenn die Erhaltung der Textinformation des Objektes im Vordergrund steht. Bei der Archivierung von Textformaten sind grundsätzliche technische Abhängigkeiten zu beachten.

- Abhängigkeit 1: der Zeichensatz (Character Set)

Einen Zeichensatz kann man sich als Tabelle vorstellen, in der ein numerischer einem Zeicheninhalt zugeordnet wird. Die Maschine nimmt den Wert der Zahl und sieht in der Zeichensatztabelle an der entsprechenden Stelle nach, in welches Zeichen die Zahl decodiert werden muss. Dieser Vorgang hat noch nichts mit der Darstellung eines Zeichens auf dem Bildschirm oder beim Druckvorgang zu tun.[5]

---

4 für eine kurze Übersicht über Bildformate s. Rohde-Enslin, Stefan (2004): *nestor - Ratgeber - Nicht von Dauer: Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen*. Berlin: nestor, IfM . S. 12ff : urn:nbn:de:0008-20041103017

5 für eine gelungene Einführung in das Gebiet der Zeichensätze s. Constable, Peter (2001):

Beispiel: Beim ASCII Zeichencode entspricht der Wert 65 (binär 01000001) dem Zeichen „A".

- Abhängigkeit 2: Schriften (Font)

Fonts geben den Zeichen eine Gestallt auf dem Bildschirm oder beim Druck. Dem Zeichen eines Zeichensatzes ist innerhalb eines Fonts ein Bild (oder mehrere Bilder) zugeordnet. Bekannte Schrifttypen sind z.B. Arial, Times New Roman usw.

Die korrekte Darstellung eines Textes ergibt sich demnach aus einer Kette von Abhängigkeiten. Um ein Textdokument mitsamt dem Schriftbild (d.h. Formatierungen, Absätze und Font) erhalten zu können, benötigt ein Archiv den korrekten Zeichensatz und den korrekten Font. Dies ist ein Problem für den dauerhaften Erhalt, denn die meisten Dateiformate, die im Bereich der Textverarbeitung verwendet werden, sind von Zeichensätzen und Fonts abhängig, die außerhalb der Textdatei gespeichert werden. Insbesondere die Zeichensätze sind oft ein Teil des Betriebssystems. Das Textverarbeitungsprogramm leistet die Verknüpfung von Code-Zeichen-Schriftzeichen und sorgt für die korrekte Darstellung des Textdokuments.

Konsequenzen für das Archiv

Für die langfristige Darstellbarkeit eines Textes muss das Archiv zumindest den oder die verwendeten Zeichensätze kennen. Die Informationen über die Zeichensätze sollten mit Bezug auf die jeweiligen Dateien in den Metadaten des Archivs fest verzeichnet sein.

Bei Neuzugängen sollte das Archiv Dateiformate wählen, die weit verbreitete und standardisierte Character Sets unterstützen. Der älteste (seit 1963) Zeichensatz ASCII kann beinahe auf allen Plattformen decodiert und dargestellt werden. Leider wurde dieser Zeichensatz allein für den amerikanischen Raum entwickelt, so dass er keinerlei Umlaute und kein „ß" enthält. Damit ist ASCII für deutsche Textdokumente nicht ausreichend. Für Archive besonders zu empfehlen sind Dateiformate, die Unicode[6], speziell UTF-8 (Unicode encoding Form neben UTF-16 und UTF-32) unterstützen. UTF-8 ist auch der empfoh-

Character set encoding basics. Understanding character set encodings and legacy encodings. In: Implementing Writing Systems: An introduction. 13.06.2001. <http://scripts.sil.org/IWS-Chapter03> (Abrufdatum: 12.12.2007)

6 Whistler, Ken/ Davis, Mark/ Freytag, Asmus (2004): *Unicode Technical Report #17. Character Encoding Model. Revision 5*. In: Unicode Technical Reports. 09.09.2004. < http://www.unicode.org/reports/tr17/> (Abrufdatum: 12.12.2007)

lene Zeichensatz für Dokumente im HTML, XML oder SGML-Format. Weit verbreitet und für Archive geeignet ist der Zeichensatz „Latin-1, Westeuropäisch“ ISO 8859-1, der auch ASCII-Texte darstellen kann.

Die gewissenhafte Dokumentation der verwendeten Zeichensätze sollte ein Archiv zumindest vor dem Verlust der reinen Textinformation bewahren. Um auch die ursprüngliche optische Form zu erhalten, sollten die technischen Informationen über die verwendeten Schriftsätze (Fonts) ebenso vom Archiv in den Metadaten nachgewiesen werden.

Bei bereits bestehenden Beständen an Textdokumenten, sollte mit geeigneten technischen Werkzeugen der zugrundeliegende Zeichensatz ermittelt werden. Sollte der ermittelte Zeichensatz nicht den oben erwähnten weit verbreiteten Standards entsprechen, empfiehlt sich auf lange Sicht wahrscheinlich eine Migration, vorrausgesetzt die geeigneten technischen Werkzeuge sind im Archiv vorhanden.

Besonders geeignete Dateiformate für Archive

Da das Archiv alle Informationen über die verwendeten Zeichensätze und Fonts sammeln und erschließen muss, sollten nur Dateiformate verwendet werden, aus denen diese Informationen auch gewonnen werden können. Dies ist bei Dateiformaten der Fall, wenn ihr technischer Aufbau öffentlich (entweder durch Normung oder Open Source) beschrieben ist. Ein Archiv sollte Textformate meiden, deren technischer Aufbau nicht veröffentlicht wurde (proprietäre Formate), da dann der Zugriff auf die für die Langzeitarchivierung wichtigen technischen Informationen kompliziert ist.

Ein Beispiel für ein offenes Dokumentformat ist das „Open Document Format“ (ODF). Der gesamte Aufbau einer ODF-Datei ist öffentlich dokumentiert. Eine Datei besteht im wesentlichen aus mehreren komprimierten XML-Dateien, die alle mit dem Zeichensatz UTF-8 gespeichert wurden. Die von ODF-Dateien verwendeten Schriftsätze sind kompatibel zu UTF-8 und in den XML-Dateien angegeben. Sollte eine ODF-Textdatei im Archiv mit den vorhandenen technischen Mitteln nicht mehr darstellbar sein, dann kann mindestens der Textinhalt und die Struktur des Dokuments über die zugrundeliegenden XML-Dateien zurückgewonnen werden.

Ein Textformat, das speziell für die Archivierung entwickelt wurde, ist das PDF/A-Format. Das Dateiformat wurde so konzipiert, dass Zeichensatz und

die verwendeten Fonds in der jeweiligen Datei gespeichert werden. Ein Textdokument im PDF/A Format ist somit unabhängiger von der jeweiligen Plattform, auf der es dargestellt werden soll.

15.2 Bilddokumente

Markus Enders

Seitdem Anfang der 1990er Jahre Flachbettscanner nach und nach in die Büros und seit Ende der 1990er Jahre auch zunehmend in die Privathaushalte einzogen, hat sich die Anzahl digitaler Bilder vervielfacht. Diese Entwicklung setzte sich mit dem Aufkommen digitaler Fotoapparate fort und führte spätestens seit der Integration kleiner Kameramodule in Mobiltelefone und Organizer sowie entsprechender Consumer-Digitalkameras zu einem Massenmarkt.
Heute ist es für Privatleute in fast allen Situationen möglich, digitale Images zu erzeugen und diese zu verschiedenen Zwecken weiterzubearbeiten. Der Markt bietet unterschiedliche Geräte zu unterschiedlichen Zwecken an: von kleinen Kompaktkameras bis zu hochwertigen Scanbacks werden unterschiedliche Qualitätsbedürfnisse befriedigt.
Entsprechend haben sich auch Softwarehersteller auf diesen Markt eingestellt. Um Bilddokumente nicht im Dateisystem eines Rechners verwalten zu müssen, existieren heute unterschiedliche Bildverwaltungsprogramme für Einsteiger bis hin zum Profifotografen.
Diese Entwicklung kommt auch den Gedächnisorganisationen zu gute. Vergleichsweise günstige Preise ermöglichen es ihnen, ihre alten, analogen Materialien mittels spezieller Gerätschaften wie bspw. Scanbacks, Buch- oder Microfilmscannern zu digitalisieren und als digitales Image zu speichern. Auch wenn Texterfassungsverfahren über die Jahre besser geworden sind, so gilt die Authentizität eines Images immer noch als höher, da Erkennungs- und Erfassungsfehler weitestgehend ausgeschlossen werden können. Das Image gilt somit als „Digitales Master", von dem aus Derivate für Online-Präsentation oder Druck erstellt werden können oder deren Inhalt bspw. durch Texterkennung / Abschreiben für Suchmaschinen aufbereitet werden kann.

Datenformate

Die seit über zwei Jahrzehnten statt findende Digitalisierung von Bildmaterialien hat zu einer Vielzahl unterschiedlicher Datenformate geführt. Gerade zu Beginn waren die technischen Faktoren limitierend, was aus Gründen schneller Implementierbarkeit und einfachen Handlings während des Betriebs zu „einfachen" technischen Lösungen führte. Diese waren teilweise so proprietär, dass sie nur von der Herstellersoftware gelesen und geschrieben werden konnten. Datenaustausch stand zu Beginn der Digitalisierung nicht im Vordergrund, sodass nur ein Teil der Daten zu Austauschzwecken in allgemein anerkannte und

unterstützte Formate konvertiert wurden.
Heute ermöglicht das Internet einen Informationsaustausch, der ohne standardisierte Formate gar nicht denkbar wäre. Der Begriff „Standard“ ist aus Sicht der Gedächnisorganisationen jedoch kritisch zu beurteilen, da „Standards“ häufig lediglich so genannte “De-facto“-Standards sind, die nicht von offiziellen Standarisierungsgremien erarbeitet und anerkannt wurden. Ferner können derartige Standards bzw. deren Unterstützung durch Hard- und Softwarehersteller lediglich eine kurze Lebenserwartung haben. Neue Forschungsergebnisse können schnell in neue Produkte und damit auch in neue Datenformate umgesetzt werden.
Für den Bereich der Bilddokumente sei hier die Ablösung des GIF-Formats durch PNG (Portable Network Graphics) beispielhaft genannt. Bis weit in die 1990er Jahre hinein war GIF der wesentliche Standard, um Grafiken im Internet zu übertragen und auf Servern zu speichern. Dieser wurde aufgrund leistungsfähigerer Hardware, sowie rechtlicher Probleme durch das JPEG- und PNG-Format abgelöst. Heute wird das GIF-Format noch weitestgehend unterstützt, allerdings werden immer weniger Daten in diesem Format generiert. Eine Einstellung der GIF-Format-Unterstützung durch die Softwarehersteller ist damit nur noch eine Frage der Zeit.
Ferner können neue Forschungsansätze und Algorithmen zu neuen Datenformaten führen. Forschungsergebnisse in dem Bereich der Wavelet-Komprimierung sowie die Verfügbarkeit schnellerer Hardware führten bspw. zu der Erarbeitung und Implementierung des JPEG2000 Standards, der wesentlich bessere Komprimierungsraten bei besserer Qualität liefert als sein Vorgänger und zeigt, dass heute auch hohe Komprimierungsraten bei verlustfreier Komprimierung erreicht werden können.
Verlustfrei ist ein Komprimierungsverfahren immer dann, wenn sich aus dem komprimierten Datenstrom die Quelldatei bitgenau rekonstruieren lässt. Verlustbehaftete Komprimierungsverfahren dagegen können die Bildinformationen lediglich annäherungsweise identisch wiedergeben, wobei für das menschliche Auge Unterschiede kaum oder, je nach Anwendung, überhaupt nicht sichtbar sind.
Neben dem oben erwähnten JPEG- oder PNG-Format, findet heute vor allem das TIFF-Format für die Master-Images Einsatz. Dessen Spezifikation beschreibt allerdings lediglich den prinzipiellen Aufbau dieses Formats und lässt viel Raum für eigene Erweiterungen und Komprimierungsmethoden. Daher lohnt sich ein genauer Blick darauf, welche Komprimierungsmethoden von einer Software unterstützt werden und welche Risiken mit deren Nutzung verbunden sind. So ist bspw. die LZW-Komprimierung für TIFF Images nach

Bekannt werden des entsprechenden Patents auf den Komprimierungsalgorithmus aus vielen Softwareprodukten verschwunden. Als Folge daraus lassen sich LZW-komprimierte TIFF Images nicht mit jeder Software einlesen, die TIFF unterstützt.
Aus Sicht der Langzeitarchivierung ist daher heute Stand der Technik die Nutzung des unkomprimierten TIFF-Formats für Graustufen- und Farbimages. Dies ist jedoch aufgrund des Platzbedarfs gerade für hochaufgelöste Images recht umstritten. Als Nachfolger wird derzeit der JPEG2000-Standard gehandelt, der vor allem in seiner verlustfreien Variante, dieselbe Qualität erreicht, jedoch wesentlich weniger Speicherplatz einnimmt. Derzeit behindert die mangelnde Unterstützung seitens der Softwarehersteller die Einsatzfähigkeit des neuen Formates: viele Programme können JPEG2000 nicht lesen oder schreiben, obwohl mittlerweile kostengünstige Programmierlibraries sowie kleine Konvertierungstools auf dem Markt sind.
Für reine schwarz-weiß (bitonale) Images hat sich die FaxG4-Komprimierung bewährt, da sie sehr gute Komprimierungsraten erlaubt und verlustfrei arbeitet.
Den oben genannten Dateiformaten ist gemein, dass sie von der Aufnahmequelle generiert werden müssen. Digitalkameras jedoch arbeiten intern mit einer eigenen an den CCD-Sensor angelehnten Datenstruktur. Dieser CCD-Sensor erkennt die einzelnen Farben in unterschiedlichen Sub-Pixeln, die nebeneinander liegen, wobei jedes dieser Sub-Pixel für eine andere Farbe zuständig ist. Um ein Image in einem gängigen Rasterimageformat generieren zu können, müssen diese Information aus den Sub-Pixeln zusammengeführt werden – d.h. entsprechende Farb-/Helligkeitswerte werden interpoliert. Je nach Aufbau und Form des CCD-Sensors finden unterschiedliche Algorithmen zur Berechnung des Rasterimages Anwendung. An dieser Stelle können aufgrund der unterschiedlichen Strukturen bereits bei einer Konvertierung in das Zielformat Qualitätsverluste entstehen. Daher geben hochwertige Digitalkameras in aller Regel das sogenannte „RAW-Format" aus, welches von vielen Fotografen als das Master-Imageformat betrachtet und somit archiviert wird. Dieses so genannte „Format" ist jedoch keinesfalls standardisiert[7]. Vielmehr hat jeder Kamerahersteller ein eigenes RAW-Format definiert. Für Gedächnisinstitutionen ist diese Art der Imagedaten gerade über längere Zeiträume derzeit nur schwer zu archivieren. Daher wird zumeist auch immer eine TIFF- oder JPEG2000 Datei zusätzlich zu den RAW-Daten gespeichert.
Die Wahl eines passenden Dateiformats für die Images ist, gerade im Rahmen

7 Zu den Standarisierungsbestrebungen siehe http://www.openraw.org/info sowie http://www.adobe.com/products/dng/

der Langzeitarchivierung, also relativ schwierig. Es muss damit gerechnet werden, dass Formate permanent auf ihre Aktualität, d.h. auf ihre Unterstützung durch Softwareprodukte, sowie auf ihre tatsächliche Nutzung hin überprüft werden müssen. Es kann davon ausgegangen werden, dass Imagedaten von Zeit zu Zeit in neue Formate überführt werden müssen, wobei unter Umständen auch ein Qualitätsverlust in Kauf genommen werden muss.

Bedeutung der Metadaten für die Archivierung

Die Speicherung und Lagerung der Imagedaten über längere Zeiträume kann dazu führen, dass Daten nur noch teilweise lesbar sind. Ebenfalls können Daten durch fehlerhafte Konvertierungsprozesse zerstört werden. Die Probleme, die zu bewältigen sind, sind vielfältig:
Während der Lagerung und Konvertierung von Daten ist der Kontext einzelner Images beizubehalten. D.h. die Zugehörigkeit einzelner Seiten oder anderer digitalisierter Objekte zu einem größeren Kontext (bspw. eines Buches) muss sichergestellt werden. Dazu ist es gerade für die Langzeitarchivierung ratsam, entsprechende Daten zusätzlich zu externen Metadatensätzen auch direkt im jeweiligen Image unterzubringen. Das TIFF-Dateiformat kennt dazu bzw. die TIFF-Tags PAGENAME, DOCUMENTNAME und IMAGEDESCRIPTION, um Informationen zu dem jeweiligen Image zu speichern. Da es sich um freie Textfelder handelt, ist prinzipiell das Abspeichern von XML-Strukturen innerhalb des Feldes möglich.

- PAGENAME kann bspw. die jeweilige Seitennummer innerhalb des Buches enthalten. Auch wenn bspw. der Dateiname eines Images verloren geht, kann immer die Reihenfolge der verschiedenen Imagedateien innerhalb des übergeordneten Kontexts bestimmt werden.
- DOCUMENTNAME sollte Informationen zum übergeordneten Kontext enthalten, die diesen eindeutig identifizieren. Dies kann der Titel, der Autor oder aber auch der Identifier (bspw. die ISBN oder eine Katalognummer) sein.
- IMAGEDESCRIPTION kann weiterführende Informationen zum Kontext des Images enthalten, bspw. die komplette bibliographische Information.

Für die Langzeitarchivierung sind auch Metadaten hinsichtlich der Generierung sowie des Generierungsprozesses wichtig. Informationen zur eingesetzten Hard- und Softwareumgebung hilfreich sein, um später bestimmte Gruppen zur Bearbeitung bzw. Migration (Formatkonvertierungen) auswählen zu können.
Im klassischen Sinn werden Formatmigrationen zwar anhand des Dateiformats

ausgewählt. Da jedoch Software selten fehlerfrei arbeitet, muss bereits bei der Vorbereitung der Imagedaten Vorsorge getroffen werden entsprechende Dateigruppen einfach selektieren zu können, um später bspw. automatische Korrekturalgorithmen oder spezielle Konvertierungen durchführen zu können.
Ein nachvollziehbares und in der Vergangenheit real aufgetretenes Szenario ist bspw. die Produktion fehlerhafter PDF-Dateien auf Basis von Images mittels einer defekten Programmbibliothek. Diese so genannten „Libraries" werden von verschiedenen Softwareherstellern häufig nur zugekauft, sodass deren Interna ihnen unbekannt sind. Tritt in dieser Programmbibliothek ein Fehler auf, so ist dieser eventuell für den Programmierer nicht auffindbar, da er seine selbst erzeugten Dateien nicht wieder einliest. Dies gilt vor allem für klassische Exportfunktionen.
In dem oben erwähnten Szenario erzeugte die entsprechende Programmbibliothek nur unter dem Solaris Betriebssystem fehlerhafte PDF-Dateien, bei denen ein „." (Punkt) durch ein „," ersetzt wurde. Kritisch für die Langzeitarchivierung wird der Fall dann, wenn einige Softwareprodukte solche Daten unbeanstandet laden und anzeigen, wie in diesem Fall der Adobe PDF-Reader. „Schwierigkeiten" machten dagegen OpenSource Programme wie Ghostscript sowie die eingebauten Postscript-Interpreter einiger getesteter Laserdrucker.
Letztlich kann dies dazu führen, dass solche Daten über Monate oder Jahre hinweg produziert werden. Werden entsprechende Informationen zur technischen Laufzeitumgebung zu jedem einzelnen Image gespeichert, kann das Data-Management eines Langzeitarchivierungssystem entsprechende Dateien identifizieren und für eine Fehlerbehebung selektieren.
Die besondere Gefahr bei der Be- und Verarbeitung dieser so genannten „Embedded"-Metadaten besteht darin, dass sie, obwohl von Standards vorgesehen häufig nicht durch entsprechende Implementierungen berücksichtigt werden. D.h. diese Metadaten gehen häufig beim Speichern nach einem Bearbeitungsschritt verloren. Für die Langzeitarchivierung bedeutet dies, dass diese Metadaten direkt vor dem Einspielen in das Langzeitarchivierungssystem überprüft und ggf. erzeugt werden müssen.

Technische Metadaten für Imagedateien

Jede Datei hat aufgrund ihrer Existenz technische Metadaten. Dies sind so genannte formatunabhängige Metadaten, die u.a. auch dazu dienen können die Authentizität eines Images zu beurteilen. Checksummen sowie Größeninformationen können Hinweise darauf geben, ob ein Image im Langzeitarchiv modifiziert wurde.
Darüber hinaus gibt es formatspezifische Metadaten. Diese hängen direkt vom

eingesetzten Dateiformat ab und enthalten bspw. allgemeine Informationen über ein Image:

- Bildgröße in Pixel und Farbtiefe
- Information über das Subformat – also bspw. Informationen zum angewandten Komprimierungsalgorithmus, damit der Datenstrom auch wieder entpackt und angezeigt werden kann.

Diese Daten lassen sich direkt aus einer Imagedatei gewinnen. Das Tool JHOVE ist bspw. in der Lage diese Daten zu erzeugen und als XML-Datei auszugeben. Im Rahmen der Langzeitarchivierung können diese Informationen sinnvoll bspw. zur Selektion von Daten verwendet werden, indem Migrationsprozesse abhängig von der jeweiligen Farbtiefe andere Zielformate definieren.

Generierungsprozess von Images

Um Images zu Einheiten zu gruppieren und diese entsprechend mit Metadaten zu beschreiben, ist der Einsatz von so genannten Containerformaten sinnvoll. Diese beschreiben ein komplexes Objekt, welches durch ein oder mehrere Images wiedergegeben wird. Diese Daten sind nicht innerhalb der Images gespeichert, sondern liegen teilweise redundant außerhalb des digitalen Objekts vor.

Ein entsprechendes Containerformat, das ein Archivsystem zum Einspielen der Images benötigt, könnte bspw. METS oder MPEG-21 DIDL sein.

Die Information zum Kontext sowie die entsprechenden Metadaten können selten sinnvoll in einem Arbeitsschritt erfasst werden. Vielmehr ist der Einsatz spezieller Software zur Steuerung von Geschäftsprozessen sinnvoll, die diese Daten erfasst, den einzelnen Images zuordnet und anschließend zusammen als ein Paket mit den Images an das Langzeitarchivierungssystem überführt. Werden Informationen zu einzelnen Arbeitsschritten erfasst, ist nachträglich auch die Beurteilung der Imagequalität im Langzeitarchiv möglich, da bspw. entsprechende Be- und Verarbeitungsmaßnahmen Rückschlüsse auf die ursprünglich erzeugte Imagedatei zulassen.

Ausblick auf die Aufbereitung von Imagedaten zur Langzeitarchivierung

Da es heute für die Generierung und Speicherung von Imagedaten bewährte Technologien gibt, hängt die Möglichkeit Bilddokumente langfristig zu archivieren und in einer ihrem Originalzustand entsprechenden oder weitgehend angenäherten Qualität verfügbar zu machen, von der Berücksichtigung der o. g. Faktoren ab.

Generell lässt sich sagen, dass eine genaue Planung und Dokumentation hinsichtlich eingesetzter Software, benutzter Formate und erfasster Metadaten diese Aufgabe vereinfachen wird. Ferner wird zukünftig Software zur Verwaltung und Steuerung von Geschäftsprozessen, gerade bei der Generierung von Bilddokumenten, die Kosten zur Erfassung dieser zusätzlichen Informationen senken. Nicht zuletzt deswegen ist zu hoffen, dass die Kosten für die Langzeitarchivierung von Bilddokumente sinken, auch wenn deren Produktionskosten zunächst leicht ansteigen werden.

15.3 Multimedia/Komplexe Applikationen

Winfried Bergmeyer

Bis zum Beginn des 20. Jahrhunderts bestanden die kulturellen Erzeugnisse, die ihren Weg in Bibliotheken, Archive und Museen fanden, in der Regel aus Büchern, Zeichnungen, Gemälden und anderen Medien, deren Nutzung ohne technische Hilfsmittel erfolgen konnte. Mit Erfindung der Fotografie, des Films und der Tonaufzeichnung hat sich das Spektrum der kulturellen Produktion neue Medien erschlossen, die das Kulturschaffen bzw. dessen Aufzeichnung revolutionierten, dabei aber technische Hilfsmittel in Form von Tonbandgeräten oder Schallplattenspielern für deren Nutzung erforderlich machten. Zum Ende des ausgehenden 20. Jahrhunderts erlebten wir mit der Revolution der Informationstechnologie eine weitere, tief greifende Veränderung. Nicht nur, dass mit dem Internet und dem Aufkommen multimedialer Anwendungen neuartige Kommunikations- und Ausdrucksformen entstanden, auch wurden und werden analoge Objekte zum Zweck der Langzeitbewahrung und der Langzeitverfügbarkeit in das digitale Format überführt. Diese digitalen Objekte sind ohne entsprechende Interpretation der Bitströme durch Computer nicht nutzbar und damit verloren. Der Auftrag zur Bewahrung des kulturellen Erbes erfordert angesichts dieser Abhängigkeiten neue Konzepte für deren Sicherung und Nutzbarkeit in Bibliotheken, Archiven und Museen.

Der Begriff „Multimedia“ bedarf in diesem Zusammenhang einer genaueren Erklärung[8]. Eigentlich beinhalten multimediale Objekte zumindest zwei unterschiedliche Medien, z. B. Ton- und Bildfolgen. Mittlerweile hat sich dieser Begriff allerdings für die Bezeichnung von Objekten mit nichttextuellen Inhalten eingebürgert. Wir werden den Begriff hier in dieser letztgenannten Form verwenden.

Vor allem im Audio- und Videobereich steht die technische Entwicklung in Abhängigkeit mit der permanenten Erschließung neuer kommerzieller Märkte. Damit ergibt sich das Problem der Obsoleszens von Hardware, Software und Dateiformaten, angeschoben durch den Innovationsdruck des Marktes. Ein Blick auf den Bereich der Tonaufzeichnung zeigt im z. B. im Hardwarebereich seit den Wachszylindern ein vielfältiges Entwicklungsspektrum über Schallplatte, Tonband, Kassette, Diskette, CD-ROM und DVD, deren Innovationszyklen sich sogar beschleunigen. Keines der Medien ist rückwärts kompatibel und ein

8 Das Wort „Multimedia“ wurde 1995 durch die Gesellschaft für deutsche Sprache zum „Wort des Jahres“ erklärt. 1995 stand der Begriff vor allem für die interaktiven Innovationen im Bereich der Computertechnologie.

Ende der technischen Fort- und Neuentwicklung ist nicht in Sicht. Dies erfordert für die kulturbewahrenden Institutionen erhebliche finanzielle, technische und personelle Anstrengungen. In der Bestandserhaltung rücken die inhaltserhaltenden Maßnahmen, beschleunigt durch den Trend zur digitalen Herstellung von Publikationen, sowie zur Digitalisierung von analogem Material, immer stärker in den Mittelpunkt[9].

Mit den sich verändernden Distributionsformen (Video-on-Demand, Filesharing u. a.) entstehen zudem neue Notwendigkeiten für die Sicherung der Urheber- und Verwertungsrechte in Form des „Digital Rights Management" mit Nutzungslimitierungen, die weitreichende Folgen für die Langzeitarchivierung, vor allem im Bereich der audiovisuellen Medien, mit sich bringen.

Besondere Anforderungen sind an die Erfassung der deskriptiven, technischen und administrativen Metadaten zu stellen. An dieser Stelle sollen nicht die verschiedenen Metadatensysteme aufgezählt werden[10], dennoch sei darauf hingewiesen, dass für die nichttextuellen Medien hier ein größerer Dokumentationsbedarf vorhanden ist. Allein die Benennung der (Container-)Formatspezifikationen ist sehr umfangreich, z. B. bei Rastergrafiken Farbrauminformationen oder Kompressionsverfahren und deren Einstellungen.

Mit dem Konzept des Universal Virtual Computer (UVC) ist ein Ansatz vorhanden, die Komplexität der Erhaltung und Erfassung von digitalen Medienobjekten zu vereinfachen[11]. Im Prinzip werden dabei die Objekten in schematisierter Form in XML mit entsprechender DTD aufgenommen ohne dabei die Objekte beständig in neue Formate zu migrieren. Der Kerngedanke ist die Entwicklung eines virtuellen Computers, der in jeder Umgebung lauffähig ist und die jeweils notwendigen Programme zur Umsetzung der Mediendaten emuliert[12]. Das Konzept des UVC wird allerdings wohl nicht für komplexe Applikationen durchführbar sein, da hier der Emulationsaufwand sehr hoch werden wird. Da zur Umsetzung ein großer Programmieraufwand geleistet werden

9 Royan, Bruce und Cremer, Monika: Richtlinien für Audiovisuelle und Multimedia-Materialien in Bibliotheken und anderen Institutionen, IFLA Professional Reports No. 85, http://www.ifla.org/VII/s35/index.htm#Projects (21.12.2007).

10 Hingewiesen sei u. a. auf MPEG-7 als „Multimedia Content Decription Interface" hingewiesen werden, das bereits von zahlreichen Institutionen verwendet wird. Siehe dazu: http://www.chiariglione.org /mpeg/standards /mpeg-7/mpeg-7.htm (1.2.2008)

11 Lorie, R. A.: Lang-Tem Archiving of Digital Information. Techn. Bericht RJ-10185 (95059). San Jose, CA: IBM Almaden Research Center, 2000.

12 Beispielhaft wurde dies an der Königlichen Bibliothek der Niederlande für Bildobjekte umgesetzt. Siehe dazu: http://www.kb.nl/hrd/dd/dd_onderzoek/uvc_voor_images-en.html (3.2.2008)

müßte, ist abzuwarten, ob und in welcher Form sich dieses Konzept etabliert.
Unter einer komplexen Applikation wird eine Datei oder eine Gruppe von Dateien bezeichnet, die als Computerprogramm ausgeführt werden können. Dies kann ein Computerspiel ebenso wie eine eLearning-Anwendung sein. Multimediale Elemente sind oftmals Bestandteil dieser Applikationen. Anders als bei den oben besprochenen, nichtexttuellen Objekten ist bei den Applikationen zusätzlich eine direkte Abhängigkeit der Nutzbarkeit vom Betriebssystem gegeben. Erst die diesen Applikationen inhärenten Programmabläufe inklusive der Einbettung multimedialer Elemente erfüllen die intendierten Aufgaben und Ziele. Diese verlangen andere Langzeitarchivierungsstrategien in Form der Emulation[13], Migration oder aber der „Technology preservation", der Archivierung der Hardware und Betriebssysteme.
Eine in diesem Zusammenhang immer wieder gestellte Frage ist die nach der Zulässigkeit dieser Emulations- und Migrationskonzepte hinsichtlich künstlerischer Werke und deren Authentizität[14]. Die zunehmenden Interaktions- und Modifikationsmöglichkeiten durch den Rezipienten, die Einfluss auf das künstlerische „Objekt" (Anwendung) haben und haben sollen, werfen zusätzliche Fragen auf, die im Rahmen der Langzeitarchivierung und der Langzeitverfügbarkeit beantwortet werden müssen[15].
Gerade am Beispiel interaktiver, multimedialer Kunst-Installationen wird die Komplexität der Aufgabe Langzeitarchivierung deutlich. Seit den 90er Jahren des 20. Jahrhunderts sind künstlerische Installation mit digitalen Elementen Bestandteil moderner Kunstproduktion, aber erst zu Beginn des neuen Jahrtausends entwickelte sich ein breiteres Problembewusstsein für die Problematik ihrer Erhaltung. Anders als bei technischen oder wissenschaftlichen Applikationen, deren Essenz in der Regel klar beschreibbar und somit überprüfbar ist, stellt sich bei den Kunstobjekten die Frage nach der Wirkung, der Rezeption,

13 Rothenberg, Jeff: Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, http://www.clir.org/PUBS/reports/rothenberg/contents.html (2.12.2007). Er fordert die Einbindung digitaler Informationen in die Emulatoren, die es ermöglichen, originäre Abspielumgebungen zu rekonstruieren.

14 Als Beispiel siehe die Diskussion um das Projekt "The Earlking". Rothenberg , Jeff: Grahame Weinbren and Roberta Friedman,The Erl King, 1982–85, in: Depocas ,Alain; Ippolito, Jon; Jones, Caitlin (Hrsg.): The Variable Media Approach - permanence through change, New York 2003, S. 101 – 107. Ders.: Renewing The Erl King, January 2006, in: http://bampfa.berkeley.edu/about/ErlKingReport.pdf (31.11.2007)

15 Rinehart, Richard: The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Centura, http://switch.sjsu.edu/web/v6n1/article_a.htm (31.11.2007)

die allein vom Künstler zu definieren ist. Oftmals gehen Raum, Material, Klang und Objekte eine Beziehung ein, die auch zeitgesteuert sein kann (Time-based Media) und die nur in ihrer Vollständigkeit das Kunstwerk definieren.
Die Obsoleszens von Software aber vor allem auch von Hardware wird nun zum essentiellen Problem. Hier muss in einem Gespräch mit dem Künstler ausgelotet werden, welche Objekte und Medien, die dieser Gefahr ausgesetzt sind, in welcher Form konserviert, migriert, emuliert oder ersetzt werden können, ohne die künstlerische Aussage zu gefährden.[16]
Können Computerprogramme durch Emulation oder Rekompilierung lauffähig gehalten werden, so wird das „Look and Feel" doch auch durch die verwendete Hardware bestimmt. Das Ziel der Bewahrung kann nicht immer allein auf die Erhaltung der digitalen Ressourcen fixiert sein, sondern kann auch die Simulation entsprechender Ausprägungen der Mensch-Maschine-Schnittstelle beinhalten. So können zwar alte Konsolenspiele (Pong, Pacman o. ä.) auf modernen Rechnern lauffähig gehalten werden, die Qualität der Konservierung ist aber dennoch abhängig von der Erhaltung oder Simulation ursprünglicher Displays und Bedienelemente (Joysticks, Gamepads, Steuerräder, Schablonen zur Sicherheitsabfrage etc.).[17]
Insbesondere für komplexe Applikationen also muss gelten, dass für die Erhaltung und Nutzungsfähigkeit beiliegendes Material in Form von Verpackungen, Handbücher, Dokumentationen etc. ebenfalls archiviert werden muss. Es ist für die weitere Nutzung der Programme notwendig die Installationsanweisungen, Programmierungsdokumentationen und Bedienungsanleitungen zu erhalten, um die Funktionsfähigkeit der Applikationen zu sichern[18]. Diese Aufgabe stellt

16 Ein Best-Practice-Beispiel findet sich z.B. im Aufsatz von Pip Laurensen der Tate Gallery zur Konservierung der Installation von Gary Hill. Laurensen, Pip: Developing Strategies for the Conservation of Installations Incorporating Time-based Media: Gary Hill's Between Cinema and a Hard Place, in: http://www.tate.org.uk /reseach/tateresearch/tatepapers/04spring/time_based_media.htm (1.2.2008)

17 Henry Lowood , Betreuer der Wissenschafts- und Technologiesammlung der Stanford University, hat hierzu mehrfach Stellung bezogen. Lowood, Henry: Playing History with Games: Steps towards Historical Archives of Computer Gaming, Presented at the Electronic Media Group Annual Meeting of the American Institute for Conservation of Historic and Artistic Works, Portland, Oregon, June 14, 2004, in: http://aic.stanford.edu/sg/emg /library/pdf/lowood/Lowood-EMG2004.pdf (1.2.2008).

18 Duranti, Luciana: Preserving Authentic Electronic Art Over The Long-Term: The InterPARES 2 Project, Presented at the Electronic Media Group, Annual Meeting of the American Institute for Conservation of Historic and Artistic Works, Portland, Oregon, June 14, 2004. Die Projekte InterPares und InterPares2 und aktuell InterPares3 setzen sich u.a. mit den Anforderungen zur Langzeitarchivierung aktueller Werke

somit erhöhte Anforderungen an die Erstellung und von umfassenden Archivierungskonzepten, z. B. auf Grundlage des OAIS (Open Archival Information System)[19].

der Bildenden und Darstellenden Künste auseinander (http://www.interpares.org/ip2/ip2_index.cfm (20.12.2007)).

19 Siehe als Beispiel der Implementierung des OAIS das Projekt „Distarnet" der Fachhochschule Basel. Melli, Markus: Distartnet. A Distributed Archival Network, http://www.distarnet.ch/distarnet.pdf (20.12.2007)) und Margulies, Simon: Distranet und das Referenzmodell OAIS, http://www.distarnet.ch/distoais.pdf (20.12.2007).

15.3.2 Audio

Winfried Bergmeyer

Die Langzeitarchivierung von Audioobjekten bildet eine Herausforderung für Bibliotheken, Archive und Museen. Ob Sprachaufnahmen, Konzerte, Tierstimmen oder Geräusche, die Variabilität der Inhalte ist groß. Das Ziel der Langzeitarchivierung ist der Erhalt dieser akustischen Informationen sowie die Sicherung ihrer Verfügbarkeit.

Die für die Speicherung auditiven Contents verwendeten Medien unterlagen in den letzten 100 Jahren einem permanenten Wandel und tun dies weiterhin. Ersten Aufzeichnungen auf Tonwalzen folgten Schellack- und Vinyl-Platten, daneben entwickelten sich die wieder beschreibbaren Medien wie Tonbänder und Kassetten unterschiedlicher Formate. Die Revolution der digitale Aufzeichnung und ihrer Wiedergabe bediente sich ebenfalls unterschiedlicher Speichermedien wie Kassetten, CDs, Minidiscs und DVDs. Mit diesem Medien- und Formatspektrum sowie den z. T. umfangreichen Datenmengen wird die Langzeitarchivierung zu einer technologischen Herausforderung. Stehen wir bei den analogen Medien vor dem Problem der physischen Zerstörung und der selten werdenden medienspezifischen Abspielgeräte, so bilden bei digitalen Daten zusätzlich die Dateiformate einen wesentlichen Aspekt der Archivierung.

Eine Speicherung auf dem gleichen Medium ist bei vielen Technologien heute kaum mehr möglich, da Speichermedien, Aufnahme- und Abspielgeräte immer weniger zur Verfügung stehen werden. Audio-Material auf älteren Tonträgern wie Walzen oder Schellackplatten wurden daher vor dem digitalen Zeitalter zur Archivierung auf Tonbändern aufgenommen. Diese, für die dauerhafte Konservierung gedachten, Tonbänder sind aber mehreren Verfallsmechanismen ausgeliefert (Entmagnetisierung, Ablösung der Trägerschichten, Sprödigkeit, Feuchtigkeitsbefall etc.) und damit stark gefährdet. Zudem gibt es weltweit zur Zeit (2007) nur noch zwei Produzenten dieser Bänder und nur noch wenige Hersteller von Abspielgeräten. Die Zukunft der Konservierung von Audio-Objekten ist die Übertragung in digitale Computerdaten. Digitale audiovisuelle Archive, wie sie von Rundfunk- und Fernsehanstalten geführt werden, sind heute so organisiert, dass sie das gesicherte Material in definierten Zeitabständen in einem neuen und damit aktuellen Format sichern. Sogenannte DMSS (digital-mass-storage-systems) beinhalten Sicherheitsmechanismen, die die Datenintegrität bei der Migration sicherstellen.

Die permanente Weiterentwicklung von Aufnahme- und Abspielgeräten sowie die Entwicklung der für die Verfügbarkeit vor allem über das Internet oder

für Mobilgeräte verwendeten Datenformate erfordert eine dauerhafte Überwachung der Technologie. Datenmigration in neue Datenformate und Speichermedien werden deshalb zum grundlegenden Konzept der Langzeitarchivierung gehören müssen. Musikarchive, die sich die Archivierung von kommerziell vertriebenen Audio-CDs zur Aufgabe gemacht haben, stellen mittlerweile bereits erste Verluste durch Zersetzung der Trägerschichten fest. Auch hier wird ein Wechsel der Speichermedien und die Migration der Daten in Zukunft nicht zu vermeiden sein.

Durch digitale Kopierschutzmechanismen versuchen die Musikverlage ihre Rechte zu sichern. Die daraus erwachsenden technischen wie auch rechtlichen Auswirkungen sind bei der Langzeitarchivierung zu berücksichtigen. Leider gibt es keine generelle Sonderreglung für Institutionen, die für den Erhalt unseres kulturelles Erbe zuständig sind. Sogenannte „Schrankenregelungen" im Urheberrechtsgesetz aus dem Jahr 2004 ermöglichen allerdings Institutionen aus kulturellen oder wissenschaftlichen Bereichen individuelle Regelungen mit den Branchenvertretern zu vereinbaren. Hier könnten auch die besonderen Aspekte für die Langzeitarchivierung geregelt werden.

Zur Digitalisierung analogen Materials benötigt man einen Analog-to-Digital-Converter (ADC), der in einfachster Form bereits in jedem handelsüblichen PC in Form der Soundkarte vorhanden ist. Professionelle Anbieter von Digitalisierungsmassnahmen verfügen allerdings über technisch anspruchsvollere Anlagen, so dass hier ein besseres Ergebnis zu erwarten ist. Es gibt mittlerweile zahlreiche Anbieter, die auch spezielle Aufnahmegeräte für die einzelnen Technologien bereitstellen, so z. B. für die Digitalisierung von Tonwalzen-Aufnahmen.

Die Qualität der Digitalisierung vorhandener analoger Objekte ist neben der Qualität des technischen Equipments vor allem von der Samplingrate und der Bit-Rate abhängig. Erstere bestimmt die Wiederholungsfrequenz, in der ein analoges Signal abgetastet wird, letztere die Detailliertheit der aufgezeichneten Informationen. Wurde lange Zeit CD-Qualität (Red Book, 44.1 kHz, 16 bit) als adäquate Archivqualität angesehen, so ist mit der technischen Entwicklung heute Audio-DVD-Qualität (bis zu 192 kHz und 24 bit) im Gebrauch. Hier sind zukünftige Weiterentwicklungen zu erwarten und für die Langzeitarchivierung zu berücksichtigen. Auf Datenkompression, die von vielen Dateiformaten unterstützt wird, sollte verzichtet werden, da es um das möglichst originäre Klangbild geht. PCM (Pulse-Code-Modulation) hat sich als Standardformat für den unkomprimierten Datenstrom etabliert. Nachbearbeitung (Denoising und andere Verfahren) zur klanglichen Verbesserung des Originals ist daher nicht vorzunehmen, da damit das originäre Klangbild verändert wird. Die nachträgliche

Fehlerbereinigung des aufgenommenen Materials ist hingegen zulässig, da es bestimmte, durch die Aufnahmetechnik bedingte, Fehlerpotentiale gibt, deren Korrektur dem Erhalt des originären Klangs dient. Bei „Born digital"-Audiodaten ist allerdings abzuwägen, ob das originale Dateiformat erhalten werden kann oder ob auf Grund der drohenden Obsoleszenz eine Format- und Medienmigration vorzunehmen ist.

In den letzten Jahren wurde die Archivierung von Audioobjekten in Form von Fileformaten zur gängigen Praxis. Als Containerfomate hat sich das WAVE-Format als de-facto-Standard durchgesetzt. Zudem findet das AIFF-Format des MacOS-Betriebssystems breite Verwendung. Beide können als stabile und langfristig nutzbare Formate gelten. Als Sonderformat für den Rundfunkbereich wurde das BWF-Format (Broadcast-Wave-Format) von der European Broadcasting Union erarbeitet. Dieses Format wird vom Technischen Komitee der International Association of Sound and Audiovisual Archives offiziell empfohlen (vgl. IASA-TC 04, 6.1.1.1 und 6.6.2.2). Das Format ist WAVE-kompatibel, beinhaltet aber zusätzliche Felder für Metadaten. Ein wesentlich ambitionierteres Formatmodell ist MPEG-21 der Moving Pictures Expert Group, das ein Framework für die Erzeugung, Produktion und die Weitergabe multimedialer Objekte bildet. Es findet bereits in einigen Audio-Archiven Anwendung.

Für die Bereitstellung des digitalen Materials zum Gebrauch können auch Formate mit verlustbehafteter Datenkompression Verwendung finden, wie dies bei der Nutzung über das Internet in Form von Streaming-Formaten (z. B. Real Audio) oder bei MP3-Format der Fall ist.

Die technischen Metadaten sollten den ganzen Digitalisierungsvorgang dokumentieren, d. h. das originale Trägermedium, sein Format und den Erhaltungszustand sowie die für seine Wiedergabe notwendigen Geräte und Einstellungs-Parameter beinhalten. Zusätzlich sind die Parameter der Digitalisierung und die verwendeten Geräte zu notieren. Für die Kontrolle des Bitstroms sind Prüfsummen zu sichern. Diese Metadaten können innerhalb der Datei (z.B. MPEG-21) oder in einer separaten Datenbank gesichert werden.

15.3.3 Langzeitarchivierung und -bereitstellung im E-Learning-Kontext

Tobias Möller-Walsdorf

Einleitung

Möchte man sich der Frage der Archivierung und Langzeitarchivierung im Kontext des E-Learnings nähern, so ist zuerst eine Differenzierung und Definition des Themenfeldes nötig, denn was konkret unter dem Begriff E-Learning verstanden wird, hat sich in den letzten Jahren stark gewandelt. Bezeichnete mit der Etablierung des Begriffs dieser in den 1990er Jahren besonders eigenständige Lern-Anwendungen, sog. Computer Based Trainings bzw. später mit der Etablierung des Internets sog. Web Based Trainings, so wird der Begriff heute allgemein weiter gefasst.

Beispielsweise definiert Michael Kerres (aktuell) E-Learning wie folgt: „Unter E-Learning (englisch electronic learning – elektronisch unterstütztes Lernen), auch E-Lernen genannt, werden alle Formen von Lernen verstanden, bei denen digitale Medien für die Präsentation und Distribution von Lernmaterialien und/oder zur Unterstützung zwischenmenschlicher Kommunikation zum Einsatz kommen."[20]

Es geht somit im E-Learning heute neben dem technisch gestützten Selbstlernen mehr auch um die Unterstützung von Präsenzlehre. Unter dem Begriff E-Learning werden daher mittlerweile eine Vielzahl unterschiedlicher Technologien zusammengefasst, deren Spektrum technisch von Autorensystemen, Simulationen, Videokonferenzen und Teleteaching, Audiomitschnitten und Podcasts, Lernmanagementsystemen bis zu Lernspielen und Web-3D-Plattformen reicht.[21] Diese Technologien können in vielen unterschiedlichen didaktischen Szenarien mit unterschiedlichstem Umfang und unterschiedlichster Ausprägung eingesetzt werden. Galt in den Anfängen E-Learning noch als Alternative zu klassischen Lernformen, so wird es heute vor allem als sinnvolle Unterstützung und Ergänzung in der Lehre und im Lernprozess eingesetzt. Das niedersächsische (Open Source-)„Erfolgsprodukt" Stud.IP ist ein gutes Beispiel für diese Entwicklung (www.studip.de). Traditionelle Lehre und E-Learning werden so gemeinsame Bestandteile eines hybriden Lernarrangements.

Dies hat zur Folge, dass bei der Betrachtung der Bereitstellung und besonders bei der Archivierung und Langzeitarchivierung das Themenfeld E-Learning in

20 http://de.wikipedia.org/wiki/E-learning 23. August 2007.

21 Vgl. http://www.elan-niedersachsen.de/index.php?id=134 23. August 2007.

zwei Bereiche aufteilen werden sollte, die differenziert betrachtet werden müssen: Gemeint ist die Unterscheidung zwischen *a) E-Learning-Kursen bzw. Kursangeboten* und *b) E-Learning-Content*. Also dem E-Learning-Kurs als organisatorische Veranstaltungsform oder virtuellen Ort der Lernorganisation und Kommunikation und E-Learning-Content als die elektronischen Materialien, die bei der Lehre und dem Lernen Einsatz finden. Hierbei kann E-Learning-Content Teil eines E-Learning-Kurses sein, es kann aber auch selbständig unabhängig von einem Kurs nutzbar sein. Ein E-Learning-Kursangebot ist auch gänzlich ohne E-Learning-Materialien möglich, beispielsweise wenn E-Learning-Komponenten wie Foren, Wikis oder elektronische Semesterapparate in einem Lernmanagementsystem eingesetzt werden.

E-Learning-Kurse

Ein großer Teil des E-Learning hat heute mit dem Einsatz neuer Medien und Technologien zur Organisation, Durchführung und Effizienzsteigerung der Lehre zu tun. Hierbei stellt sich die Frage, was von den dabei anfallenden Daten auf den Servern der Bildungseinrichtungen archiviert werden sollte? Welchen Sinn macht es E-Learning-Kurse zu archivieren bzw. welche Bestandteile eines E-Learning-Kurses sollten bzw. müssten archiviert werden: Veranstaltungsdaten, Teilnehmerlisten, Foreneinträge und Chats, Umfragen, Test- und Prüfungsergebnisse?
Da diese Informationen zu E-Learning-Kursen sehr stark personenbezogen sind, hat eine Archivierung dieser Daten eher einen reinen Archivierungscharakter und nur wenig Aspekte einer Nachnutzbarkeit und Weiterverwertung; der Zugriff auf diese Daten wäre aus Datenschutzgründen stark eingeschränkt.
Die genannten Bestandteile der E-Learning-Kurse sind technisch sehr eng mit dem System zur Kursorganisation (beispielsweise dem Lernmanagement-System) oder einem E-Learning-Tool (z.B. für Foren und Wikis) verbunden, so dass für die Archivierung zukünftig eine Emulationsumgebung des gesamten Systems (inkl. beispielsweise der Datenbank) notwendig wäre. Alternativ könnte nur ein Export einzelner, losgelöster Bestandteile des Kurses (beispielsweise die Foreneinträge in Textform) erfolgen.

E-Learning-Content

E-Learning-Content bezeichnet in dieser Aufteilung im Gegensatz zu den E-Learning-Kursen die elektronischen Lehr- und Lernmaterialien, die im E-Learning eingesetzt werden. Die Art dieses E-Learning-Contents ist sehr heterogen

und vom technischen System und didaktischen Szenario abhängig. Es kann sich u. a. um reine Textdateien, Bilddateien, Power-Point-Präsentationen, Audio- und Videodateien, Simulationen und Animationen (Flash-Dateien), HTML-Projekte und komplexe Multimedia-Programme handeln.
Oftmals handelt es sich um unterschiedlichste multimediale und dynamische Objekte, die zusätzlich durch Interaktionen mit dem Nutzer gesteuert werden, also einer komplexen Programmierung folgen. Eine Vielzahl technischer Formate, unzureichende Normierung und besonders ein sehr hoher Innovationszyklus bei den Dateiformaten der multimedialen Objekte, machen das Thema der Archivierung von E-Learning-Content zu einem der Komplexesten, vergleichbar vielleicht mit der Archivierung von Multimedia-Anwendungen oder Computerspielen.
Werden die Dateien archiviert, besteht zudem die Gefahr, dass sie – losgelöst vom Kontext und ohne den Kurszusammenhang – didaktisch unbrauchbar oder für den Lehrenden und Lernenden inhaltlich unverständlich werden. Zusätzlich können rechtliche Aspekte den zukünftigen Zugriff auf diese Archivmaterialien erschweren, da für den Einsatz im Kurs-Zusammenhang des E-Learning-Kurses andere rechtliche Rahmenbedingungen für den E-Learning-Content bestehen, als bei frei zugänglichen Materialien (§52a UrhG).
E-Lerning-Content ist oftmals in einem technischen, proprietären System erstellt bzw. bedarf eines speziellen E-Learning-Systems um ihn anzuzeigen. Beispielsweise bei Kurs-Wikis, Contentmanagement-Systemen oder speziellen Authoring-Tools wie z.B. ILIAS. Ist ein Export der Materialien in ein Standardformat möglich bzw. wurden die Materialien bereits in einem gebräuchlichen Format erstellt, so ist die Archivierung einfacher. Die möglichen Formate, die im E-Learning zum Einsatz kommen, entsprechen zum größten Teil den gebräuchlichen Multimedia-Formaten, also beispielsweise PDF, Power-Point, Flash, AV-Formate, HTML-Projekte. Dazu aber auch noch Spezialformate wie z.B. Dateien des weit verbreiteten Aufzeichnungstools Lecturnity.[22]
Um die Lesbarkeit digitaler Materialien möglichst lange zu gewährleisten, sollten allgemein Datenformate verwendet werden, deren Spezifikation offen gelegt ist (z.B. ODF, RTF, TIFF, OGG). Proprietäre Formate, die an die Produkte bestimmter Hersteller gebunden sind, wie z.B. DOC oder PPT, sind zu vermeiden. Der Grund hierfür liegt darin, dass langfristig zumindest die Wahrscheinlichkeit hoch sein sollte, dass eine Interpretationsumgebung (Hardware, Betriebssystem, Anwendungsprogramm) für das archivierte Objekt in der künftigen Nutzergemeinde vorhanden sein wird. Diese Forderung ist für den Bereich E-Learning allerdings heute nur schwer umsetzbar. Auf jedem Fall soll-

22 http://www.lecturnity.de/ 23. August 2007.

ten aber für die Erstellung von E-Learning-Content die auch in anderen Bereichen übliche Multimediaformate eingesetzt werden. Die Archivierung ist dann zumindest analog zu anderen multimedialen Objekten zu sehen, natürlich mit allen dort auftretenden Schwierigkeiten der Emulierung oder Migration.

Archivierungskriterien

Betrachtet man beispielsweise den im Rahmen des Projektes ELAN in Niedersachsen entstandenden E-Learning-Content (www.elan-niedersachsen.de), so zeigt sich, dass nicht alle entstehenden E-Learning-Materialien auch langfristig relevant sind und nicht immer eine Archivierung und Bereitstellung mit dem Zweck der Nachnutzung und Weiterverwendung sinnvoll ist. Oftmals wandeln sich Kurse pro Semester so stark, dass von der Seite der Dozenten kein Interesse an der Archivierung und späteren Bereitstellung besteht. Eine Selektion des Materials, besonders unter dem Aspekt der Nachnutzbarkeit, ist daher angebracht. Allerdings sollte bei der Archivierung die Meinung des Autors bezüglich der Relevanz der Archivierung nicht immer ausschlaggebend sein, denn für viele Materialien ist es derzeit nur sehr schwer vorhersehbar, welcher Wert ihnen in Zukunft beigemessen wird. Dass heute beispielsweise sehr frühe (Magnetophon-)Aufzeichnungen der Vorlesungen von Max Planck als großer Glücksfall angesehen werden, war zum Zeitpunkt ihrer Erstellung in vollem Umfang sicher noch nicht abschätzbar.[23] Das „absehbare historische Interesse" ist somit besonders für Bibliothekare und Archivare, die mit diesen Materialien zu tun haben, eine der wichtigen und auch schwierigen Fragen bei der Archivierung.

Auch für die Dozenten interessant ist bei der Archivierung die Wiederverwendbarkeit und Nachnutzung von Lehrmaterial. Hier sind beispielsweise Unterlagen für Grundlagen-Vorlesungen zu nennen. Material also, das in der gleichen Form regelmäßig verwendet wird und sich ggf. nur in seiner jeweiligen Zusammenstellung unterscheidet. Solche Materialien könnten zudem über die Universität hinaus im Umfeld von Weiterbildung und Erwachsenenbildung (Lifelong Learning) eingesetzt werden. Auch Kostenreduktion bei zum Teil sehr kostenintensiven E-Learning-Produktionen, wie z.B. Videoaufzeichnungen oder komplexen Multimedia-Anwendungen, könnte bei der Archivierung eine Rolle spielen (vgl. z.B. die IWF Campusmedien[24]).

Ein weiterer Grund für die Archivierung von erstellten Lehr-, Lern- und besonders Prüfungsmaterialien können zukünftig rechtliche Anforderungen sein,

23 http://webdoc.sub.gwdg.de/ebook/a/2002/nobelcd/html/fs_planck.htm 23 August 2007.

24 http://www.iwf.de/campusmedien/ 23. August 2007.

nämlich zur späteren Kontrolle von Prüfungsergebnissen. Derzeit besteht allerdings noch keine konkrete rechtliche Verpflichtung, solche E-Learning-Dokumente längerfristig zu archivieren. Bei weitergehender Etablierung von E-Learning-Bestandteilen, besonders durch den Anstieg der nötigen Prüfungsleistungen beispielsweise bei den Bachelor-Master-Studiengängen, wird sich diese Situation aller Voraussicht nach zukünftig ändern.

Metadaten für E-Learning-Kurse und E-Learning Content

Um die Bereitstellung von E-Learning-Archivobjekten, also E-Learning-Kursen und E-Learning-Content oder Bestandteile daraus zu gewährleisten, werden neben technischen Metadaten inhaltsbeschreibende Metadaten und nachhaltig gültige Identifikatoren (Persistent Identifier) für die zu archivierenden Objekte benötigt. Nur anhand dieser Metadaten ist eine Suche in den Datenbeständen möglich. Im Bereich der Metadaten erfolgt u. a. im Rahmen von ELAN eine rege Forschungsaktivität mit Fokus auf der Entwicklung von Standards für solche Metadaten. Welche inhaltsbeschreibenden Metadaten für E-Learning-Objekte geeignet sind und an welchen bestehenden Standard (z.B. Dublin Core, LOM) sie orientiert werden, wurde im Rahmen des ELAN-Projektes in Niedersachsen ausgearbeitet, auf die Ergebnisse des „ELAN Application Profile" sei hier verwiesen.[25] Daneben ist das vom Bundesministerium für Wirtschaft und Technologie (BMWi) geförderte Projekt Q.E.D. (http://www.qed-info.de) zu nennen, welches das Ziel verfolgt, die Etablierung von innovativen Lernszenarien und eben auch internationalen Qualitätsstandards und Normen im E-Learning in Deutschland weiterzuentwickeln. Projektpartner ist unter anderem das Deutsche Institut für Normung e.V (DIN).
Bei allen diesen Bemühungen der Erfassung von Metadaten und Standardisierung mit dem Ziel der strukturierten Bereitstellung, Archivierung und Langzeitarchivierung sollten die Bibliotheken und Archive mehr als bisher in die Entwicklungsprozesse eingebunden werden. E-Learning-Content sollte, wie andere elektronische Materialien auch, in den regulären Geschäftsgang besonders der Bibliotheken einfließen und damit auch unabhängig von Projekten und temporären Initiativen Berücksichtigung finden. Nur so ist eine langfristige Bereitstellung und Archivierung dieses Teils unseres kulturellen Erbes möglich.

25 DINI Schriften 6: ELAN Application Profile: Metadaten für elektronische Lehr- und Lernmaterialien [Version 1.0, Oktober 2005]. http://nbn-resolving.de/urn:nbn:de:kobv:11-10050226. August 2007.

15.3.4 Interaktive Applikationen

Dirk von Suchodoletz

Die Diskussionen und Forschung zur digitalen Langzeitarchivierung von statischen digitalen Primärobjekten, wie Dokumenten, Digitalisaten oder Audio- und Videodatenströme sind bereits recht weit vorangekommen. Anders liegt der Fall für dynamische digitale Objekte: Sie kommen in der aktuellen Langzeitarchivierungsdebatte bisher fast nicht vor. Zu ihnen zählen:

- Betriebssysteme. Sie erlauben gemeinsam mit der physikalischen Hardware eines jeden Rechners überhaupt erst seinen sinnvollen Betrieb. Ihre Aufgabe besteht in der Steuerung der Hardware, der Ressourcenverwaltung und -zuteilung. Sie erlauben die Interaktion mit dem Endanwender. Sie sind im kompilierten Zustand – übersetzt aus dem Quellcode in ausführbaren Maschinencode - deshalb nur auf einer bestimmten Architektur ablauffähig. Mit der Übersetzung erfolgte die Anpassung an bestimmte Prozessoren, die Art der Speicheraufteilung und Peripheriegeräte zur Ein- und Ausgabe. Da eine Reihe von Funktionen von verschiedenen Programmen benötigt werden, sind diese oft in sogenannte Bibliotheken ausgelagert. Programme, die nicht alle Funktionen enthalten, laden benötigte Komponenten dynamisch aus den Bibliotheken zur Laufzeit nach. Bibliotheken und Programme hängen dementsprechend eng miteinander zusammen.
- Anwendungsprogramme. Sie setzen auf der Betriebssystemebene auf. Für die schematische Darstellung der Arbeit eines Computers wird oft ein Schichtenmodell gewählt, das die Anwendungen in der obersten Ebene anzeigt (Abbildung 15.3.4.1). Applikationen sind Programme, die für bestimmte, spezialisierte Aufgaben erstellt wurden. Mit diesen Programmen generierten und bearbeiteten Endanwender Daten der verschiedensten Formate. Der Programmcode wird im Kontext des Betriebssystems ausgeführt. Er kümmert sich um die Darstellung gegenüber dem Benutzer und legt fest, wie beispielsweise die Speicherung von Objekten in einer Datei organisiert ist und der Anwender mit dem Computer interagiert. Das Betriebsystem übernimmt die Speicherung von Dateien auf Datenträgern üblicherweise angeordnet in Verzeichnissen. Zur Ausführung auf einer bestimmten Rechnerarchitektur werden Betriebssysteme und Applikationen aus dem sogenannten Quellcode in den passenden Binärcode übersetzt. Deshalb können Programme und Bibliotheken nicht beliebig zwischen verschiedenen Betriebssystemen verschoben werden.
- Computerspiele (eigener Abschnitt)

- Interaktive Medien zur Lehre und Unterhaltung (eigener Abschnitt)
- Datenbanken - zählen ebenfalls zu den sehr frühen Anwendungen von Rechnern. Die Bewegung, Durchsuchung und Verknüpfung großer Datenbestände gehört zu den großen Stärken von Computern. Diese elektronischen Datenbestände stellen oft die Grundlage für abgeleitete Objekte dar. Zur Klasse der datenbankbasierten Anwendungen zählen Planungs- und Buchhaltungssysteme, wie SAP, elektronische Fahrpläne diverser Verkehrsträger bis hin zu Content Management Systemen moderner Internet-Auftritte von Firmen und Organisationen. Wenn von einer Datenbank sehr verschiedene Ansichten ad-hoc erzeugt werden können, ist sehr schwer abzusehen, welche dieser Ansichten zu einem späteren Zeitpunkt noch einmal benötigt werden könnten. Unter Umständen hat man sich dann auf Teilmengen festgelegt, die von nachfolgenden Betrachtern als unzureichend oder irrelevant eingestuft werden. Gerade bei Datensammlungen wichtiger langlebiger Erzeugnisse wie Flugzeugen, Infrastrukturen oder Gebäuden besteht großes allgemeines Interesse eines zeitlich unbeschränkten Zugriffs.

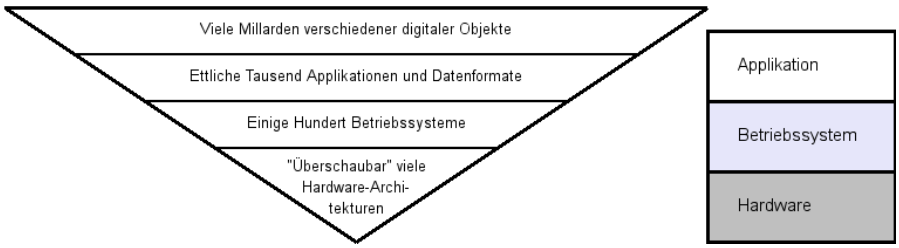


Abbildung 15.3.4.1: Sicht auf die Anzahl der nachzubildenden Objekte (links) je nach Wahl der Schicht für typische Rechnerplattformen (rechts)

Dynamische Daten zeichnen sich dadurch aus, dass sie außerhalb eines festgelegten digitalen Kontexts nicht sinnvoll interpretiert und genutzt werden können. Eine Überführung in ein analoges Medium scheidet aus: Der Ausdruck des Programmcodes auf Papier oder die Aufnahme einer Programmsitzung auf Video sind derart „verlustbehaftete" Speicherverfahren, dass sie im Sinne der Archivierung in keiner Weise die Anforderungen erfüllen.

Emulation – Erhalt von Ablaufumgebungen

Emulation setzt nicht am digitalen Objekt selbst an, sondern beschäftigt sich mit der Umgebung, die zur Erstellung dieses Objektes vorlag. Das bedeutet beispielsweise die Nachbildung von Software durch andere Software, so dass es für ein digitales Objekt im besten Fall keinen Unterschied macht, ob es durch

die emulierte oder durch die Originalumgebung behandelt wird. Ebenso kann Computerhardware durch Software nachgebildet werden, auch wenn dieses erst einmal deutlich komplexer erscheint. Generell lässt sich feststellen, dass für aktuelle Computerplattformen üblicherweise drei Ebenen für den Ansatz von Emulation identifiziert werden können:

Emulation oder Ersatz von Applikationen durch andere, um so bestimmte Datenformate darzustellen oder auszuführen. In der heutigen Softwarelandschaft ist es üblich bestimmte Applikationen oder deren Funktionalität nachzuprogrammieren, um sie in eigene Produkte zu integrieren oder den Funktionsumfang bestehender Software zu erweitern. Für viele Anwender sicherlich am naheliegendsten ist die Emulation der Eigenschaften älterer Applikationen in aktuellen Anwendungsprogrammen. Die Interpretation von einer Applikation nicht-eigenen aktuellen Datenformaten fällt ebenso in diesen Bereich. Auf diese Weise wird überhaupt ein Datenaustausch zwischen verschiedenen Programmen unterschiedlicher Hersteller erst möglich. Deshalb gehören solche Funktionen oftmals zum Standardumfang einer Applikation. Ein typisches Beispiel dieser Art von „Emulation" einer Applikation ist die Benutzung von StarOffice oder OpenOffice, um sich Microsoft Word Dokumente oder Excel Tabellen in einer Linux- oder Unix-Umgebung anzusehen. Dieses trifft auf eine ganze Reihe weiterer Applikationen zu.

Emulation von Betriebssystemen, um so auf einem gegebenen Betriebssystem Applikationen auszuführen, die die Schnittstellen (API) eines anderen Betriebssystems erwarten. Bezogen auf das Schichtenmodell (Abbildung 1) setzt diese Art der Emulation auf der Betriebssystemebene an. Die Emulation von Betriebssystemen zielt darauf ab, dass sich beispielsweise Programme unter Linux/X11 ausführen lassen, die ursprünglich für Microsoft Windows geschrieben wurden. Die Motivation liegt nicht primär in der Langzeitarchivierung begründet, sondern darin dem Zwang auszuweichen, für eine bestimmte gewünschte Applikation auch ein anderes Betriebssystem verwenden zu müssen. Besonders hervorheben kann man an dieser Stelle das Wine-Projekt oder Cygwin (Posix unter Windows).

Nachbildung eines kompletten Rechners einer bestimmten Architektur. Der Ansatz die komplette Hardware einer bestimmten Rechnerplattform zu archivieren, nutzt die tiefste Schicht (Abbildung 15.3.4.1). Deshalb reicht der Ansatz die komplette Hardware einer Computerarchitektur in Software nachzubilden deutlich weiter. Dieses scheint auf den ersten Blick sehr aufwändig, hat jedoch entscheidende Vorteile: Die Schnittstellen sind offengelegt, da oft recht verschiedene Anbieter ihre Betriebssysteme für die verschiedenen Architekturen entwickelt haben.

Emulation bedeutet also zunächst nur die Erschaffung einer virtuellen Umgebung in einer gegebenen Ablaufumgebung, üblicherweise in dem zum Zeitpunkt des Aufrufs üblichen Computersystem. Emulatoren bilden somit die Schnittstelle, eine Art Brückenfunktion, zwischen dem jeweils aktuellen Stand der Technik und einer längst nicht mehr verfügbaren Technologie. Dabei müssen sich Emulatoren um die geeignete Umsetzung der Ein- und Ausgabesteuerung und der Peripherienachbildung bemühen.

Die Auswahl der inzwischen kommerziell erhältlichen oder als Open-Source-Software verfügbaren Emulatoren oder Virtualisierer ist inzwischen recht umfangreich geworden, so dass häufig sogar mehr als ein Emulator für eine bestimmte Rechnerarchitektur zur Verfügung steht. Jedoch eignet sich nicht jeder Emulator gleichermaßen für die Zwecke des Langzeitzugriffs.
So existiert für frühe Architekturen keine deutliche Unterscheidung zwischen Betriebssystem und Applikation. Home-Computer verfügten über eine jeweils recht fest definierte Hardware, die zusammen mit einer Art Firmware verbunden ausgeliefert wurde. Diese Firmware enthält typischerweise eine einfache Kommandozeile und einen Basic-Interpreter. Nicht alle für den Betrieb von Emulatoren benötigten Komponenten, wie beispielsweise genannte Home-Computer-Firmware, ist frei verfügbar. Sie müssen ähnlich wie ein Betriebssystem für spätere Architekturen erworben worden sein. Lediglich für einige Systeme existieren frei verfügbare Nachprogrammierungen. Für die X86-Architektur beispielsweise liegt der Fall mit der Verfügbarkeit eines Open Source Systems und Grafikkarten-BIOS einfacher.
Der überwiegende Anteil von Emulatoren und Virtualisierern wurde oftmals aus ganz anderen als Langzeitarchivierungsgründen erstellt. Sie sind heutzutage Standardwerkzeuge in der Software-Entwicklung. Nichtsdestotrotz eignen sich viele dieser Werkzeuge für eine Teilmenge möglicher Langzeitarchivierungsaufgaben. Institutionen und private Nutzern reichen für zeitlich befristete Archivierung derzeitig verfügbare Programme oftmals aus.

Softwarearchiv als eine Erfolgsbedingung für Emulationsstrategien

Je nach gewählter Ebene der Emulation werden zusätzliche Softwarekomponenten benötigt. Bei der Emulation von Applikationen für den alternativen Zugriff auf ein bestimmtes Datenformat entfällt der Bedarf an weiterer Software. Optimalerweise laufen die emulierten Applikationen auf einer aktuellen Plattform und erlauben aus dieser heraus den direkten Zugriff auf die digitalen Objekte des entsprechenden Formates. Solange es gelingt die entsprechende Applikation

bei Plattformwechseln zu migrieren oder bei Bedarf neu zu erstellen, ist dieser Weg für die Langzeitarchivierung bestimmter Dateitypen durchaus attraktiv. Vorstellbar wäre dieses Verfahren für statische Dateitypen wie die verschiedenen offenen und wohldokumentierten Bildformate.
Die Emulation eines Betriebssystems oder dessen Schnittstellen erlaubt theoretisch alle Applikationen für dieses Betriebssystem ablaufen zu lassen. Dann müssen neben dem Emulator für das entsprechende Betriebssystem auch sämtliche auszuführende Applikationen in einem Softwarearchiv aufbewahrt werden (Abbildung 15.3.4.2). Bei der Portierung des Betriebssystememulators muss darauf geachtet werden, dass sämtliche in einem Softwarearchiv eingestellten Applikationen weiterhin ablaufen können.

Die Nachbildung einer kompletten Hardware in Software verspricht die besten Ergebnisse und verfolgt den generellsten Ansatz. Nun benötigt man jedoch in jedem Fall mindestens eines oder je nach Bedarf mehrere Betriebssysteme, die sich als Grundlage der darauf aufesetzenden Applikationen ausführen lassen. Das bedeutet für ein Softwarearchiv, dass neben dem Emulator für eine Plattform auch die darauf ablauffähigen Betriebssysteme aufgehoben werden müssen. Das gilt ebenfalls für die darauf basierenden Applikationen, die zur Darstellung der verschiedenen Datenformate erforderlich sind. Zusätzlich zu den benötigten Applikationen sind oft eine Reihe von Hilfsprogrammen oder -komponenten erforderlich. Teilweise sind Dateien komprimiert oder in einem gepackten Archiv zusammengeführt und müssen vor dem Zugriff erst entpackt werden. Einige Dokumente erfordern zusätzliche Schriftarten zu ihrer Darstellung, wie einige digitale Videos nur mit einem bestimmten Codec abgespielt werden können. Erfolgt eine Portierung, also Migration des Hardwareemulators, muss anschließend überprüft werden, dass die gewünschten Betriebssysteme weiterhin ablauffähig bleiben. Da die meisten Applikationen lediglich die Schnittstellen des Betriebssystems nutzen (sollten), folgt ihre Funktionsfähigkeit direkt aus der der Betriebssysteme. Betriebssysteme benötigen zur Ansteuerung der Hardware, wie Netzwerk- und Soundkarte oder Grafikadapter passende Treiber, die mit den jeweiligen Komponenten korrespondieren. Deshalb kann das äußere Update des Emulators oder Virtualisierers, neben der Migration des Containerformats der virtuellen Festplatte, auch eine Migration des Treiber-Sets des darin installierten Betriebssystems nach sich ziehen.
Dieses Dilemma löst beispielsweise VMware derzeit noch durch die Pflege und Bereitstellung geeigneter Treiber für jedes offiziell unterstützte Betriebssystem. Noch ist die Liste der Treiber sehr lang und im Sinne der Archivierung fast vollständig. Ein weiteres Problem ist die Art der Prozessor-Nutzung. Viele virtuelle

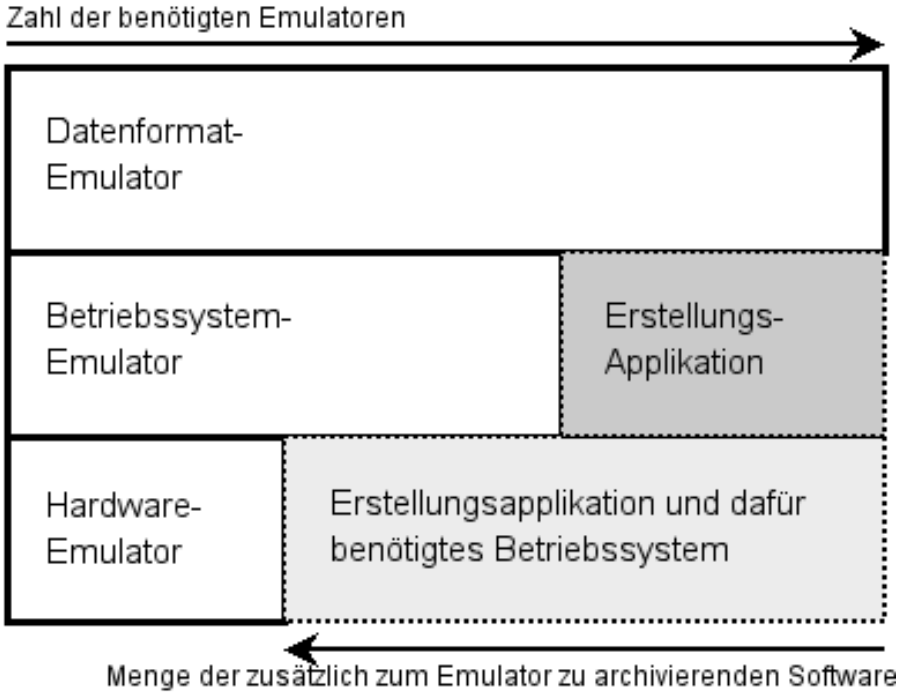


Abbildung 15.3.4.2: Je nach gewählter Ebene der Emulation für die spätere erfolgreiche Darstellung des interessierenden Primärobjekts notwendige Menge von Sekundärobjekten

Maschinen reichen CPU-Befehle des Gastes direkt an die Host-CPU weiter. Das setzt voraus, dass diese mit diesen Befehlen umgehen kann.
Geeigneter im Sinne einer wirklichen Langzeitarchivierung sind quelloffene Implementierungen, wie QEMU, Bochs oder das Java-basierte Dioscuri. Sie erlauben zum einen die Übersetzung für die jeweilige Plattform und zum anderen auch langfristige Anpassungen an völlig neue Architekturen. Zudem kann sichergestellt werden, dass auch alte Peripherie, wie ISA-Netzwerk- und Soundkarten, dauerhaft virtualisiert wird und nicht einem Produktzyklus zum Opfer fällt. Für alte Betriebssysteme, wie DOS oder Windows 3.1 besteht andernfalls die Gefahr, dass sie nur noch beschränkt ausgeführt werden können.
Ein weiterer Punkt ist die unter Umständen notwendige Anpassung der Containerdateien, in denen die Gastsysteme auf der virtuellen Maschine installiert sind. Ändert der Emulator das Datenformat, sind diese Dateien genauso wie andere digitale Objekte in ihrer Les- und Interpretierbarkeit gefährdet. Üblicherweise stellen jedoch die kommerziellen Anbieter Importfunktionen für Vorgängerversionen zur Verfügung. Bei freien, quelloffenen Emulatoren kann

alternativ zur Weiterentwicklung dafür gesorgt werden, dass ein bestimmtes Dateiformat eingefroren wird.
Ein Softwarearchiv sollte daher so angelegt sein, dass auch künftige Änderungen der Emulationsstrategie, bedingt durch veränderte Szenarien des Zugriffs auf die dynamischen digitalen Objekte, realisiert werden können.

Softwarearchiv – Komponenten

Nach den eher theoretisch angelegten Vorbetrachtungen steht nun das Softwarearchiv als Hilfsmittel der Emulationsstrategie im Mittelpunkt. Der Erhalt möglichst originalgetreuer Ablaufumgebungen für die verschiedensten Typen digitaler Objekte erfordert nicht nur die Bereitstellung der Primärwerkzeuge. Neben den eigentlichen Emulatoren sind eine ganze Reihe weiterer Software-Komponenten erforderlich.
Die Hardwareemulation setzt auf der untersten Ebene des Schichtenmodells an. Das bedeutet auf der einen Seite zwar einen sehr allgemeinen Ansatz, erfordert umgekehrt jedoch eine ganze Reihe weiterer Komponenten. Um ein gegebenes statisches digitales Objekt tatsächlich betrachten zu können oder ein dynamisches Objekt ablaufen zu sehen, müssen je nach Architektur die Ebenen zwischen der emulierten Hardware und dem Objekt selbst "überbrückt" werden. So kann ein Betrachter nicht auf einer nackten X86-Maschine ein PDF-Dokument öffnen. Er braucht hierfür mindestens ein Programm zur Betrachtung, welches seinerseits nicht direkt auf der Hardware ausgeführt wird und deren Schnittstellen direkt programmiert. Dieses Programm setzt seinerseits ein Betriebssystem als Intermediär voraus, welches sich um die Ansteuerung der Ein- und Ausgabeschnittstellen der Hardware kümmert.
Ein weiteres nicht zu unterschätzendes Problem liegt in der Form des Datenaustauschs zwischen der im Emulator ablaufenden Software und der Software auf dem aktuellen Host-System. Diese Fragestellung unterscheidet sich nicht wesentlich vom Problem des Datenaustauschs zwischen verschiedenen Rechnern und Plattformen. Mit fortschreitender technischer Entwicklung ergibt sich unter Umständen jedoch ein größer werdender Abstand zwischen dem technologischen Stand des stehenbleibenden emulierten Systems und dem die Emulation ausführenden Host-System. Zum Teil halten die verfügbaren Emulatoren bereits Werkzeuge oder Konzepte vor, um die Brücke zu schlagen.
Da sowohl die für den Anwender interessanten Daten und Anwendungsprogramme als auch die zur Ansicht notwendigen Hilfsprogramme, Emulatoren, Betriebssysteme digitale Objekte sind, sollen sie wie folgt unterschieden werden:

- Ein Zielobjekt oder Primärobjekt gehört zu den primär interessierenden

Objekten eines Archivbenutzers, welches er in irgendeiner Form geeignet betrachten will.

- Für die Betrachtung wird Software (Erstellung einer Nutzungsumgebung für das Primärobjekt) benötigt. Die notwendigen Komponenten ergeben sich aus dem jeweiligen View-Path (siehe unten) und können je nach Wahl dessen variieren. Diese Software in den verschiedenen Ausformungen wird als Hilfs- oder Sekundärobjekt bezeichnet. Es ist für den Benutzer nicht primär von Interesse, wird aber gebraucht, um überhaupt mit dem Zielobjekt umgehen zu können.
- Je nach Art des Zielobjektes können eine ganze Reihe verschiedener Sekundärobjekte notwendig sein.

Das Softwarearchiv selbst kann wieder als Bestandteil eines größeren Archivs nach dem OAIS-Modell angesehen werden. Für die Aufbewahrung der Emulatoren, der Betriebssysteme, Applikationen und Hilfsprogramme gelten die identischen Regeln, wie für die eigentlichen digitalen Primärobjekte. Trotzdem kann es von Interesse sein, diese Daten in direkt zugreifbarer Weise oder auch in speziell aufbereiteter Form vorzuhalten.

Ein weiterer Problemkreis ergibt sich aus der Art der Sekundärobjekte. Anders als bei den meisten Zielobjekten werden im Laufe der Zeit Änderungen oder Ergänzungen notwendig, die im Archiv berücksichtigt werden sollten. Da es sich hierbei um Software handelt, ist diese auf eine bestimmte Nutzungsumgebung angewiesen. Daraus folgt, dass man zum einen diese Umgebung geeignet rekonstruieren muss, um dann in dieser die gewünschten Daten anzusehen oder in selteneren Fällen zu bearbeiten. Andererseits wird sich je nach Erstellungsdatum des Objektes die damalige Erstellungs- oder Nutzungsumgebung dramatisch von der jeweils aktuellen unterscheiden. In der Zwischenzeit haben sich mit einiger Wahrscheinlichkeit die Konzepte des Datenaustausches verändert. Hier ist nun dafür zu sorgen, dass die interessierenden Primärobjekte geeignet in die (emulierte) Nutzungsumgebung gebracht werden können, dass die Betrachtung für den Archivnutzer in sinnvoller Form möglich ist und dass eventuell Bearbeitungsergebnisse aus der Nutzungsumgebung in die aktuelle Umgebung transportiert werden können.

Das Archiv muss deshalb eine ganze Reihe verschiedener Softwarekomponenten umfassen, so sind:

- die Emulatoren zu speichern, so dass mit ihrer Hilfe die Wiederherstellung einer Rechner-Architektur für bestimmte Nutzungsumgebungen erfolgen kann.
- die Betriebssysteme abzulegen, die je nach Rechnerplattform einen Teil der Nutzungsumgebung ausmachen.
- Treiber der Betriebssysteme zusätzlich zu speichern, da sie den Betriebs-

systemen überhaupt erlauben mit einer bestimmten Hardware umzugehen.
- die Applikationen zu archivieren, mit denen die verschiedenen digitalen Objekte erstellt wurden. Diese Applikationen sind ebenfalls Bestandteil der Nutzungsumgebung des Objektes. Sie sind in vielen Fällen auf die vorgenannten Betriebssysteme angewiesen.
- die unter Umständen notwendigen Erweiterungen einer Applikationsumgebung, wie bestimmte Funktionsbibliotheken, Codecs oder Schriftartenpakete zur Darstellung.
- Hilfsprogramme zu sammeln, welche den Betrieb der Emulatoren vereinfachen oder überhaupt erst ermöglichen. Hierzu zählen beispielsweise Programme, die direkt mit dem jeweiligen Containerformat eines Emulators umgehen können.
- je nach Primärobjekt oder gewünschter Nutzungsumgebung sind mehrere Varianten derselben Software zu archivieren, um beispielsweise die Lokalisierung in eine bestimmte Ein- und Ausgabesprache zu erreichen.

View-Paths und Referenzumgebungen

Digitale Objekte können nicht aus sich alleine heraus genutzt werden, sie bedürfen eines geeigneten Kontextes, damit auf sie zugegriffen werden kann. Dieser Kontext, Nutzungsumgebung genannt, muss geeignete Hardware- und Softwarekomponenten so kombinieren, dass je nach Objekttyp die Erstellungsumgebung oder ein geeignetes Äquivalent abgebildet werden.
Die Wiederherstellung von Nutzungsumgebungen oder geeigneter Äquivalente läßt sich am besten durch sogenannte „View-Paths“, Wege ausgehend vom Primärobjekt des Interesses zur Arbeitsumgebung des Betrachters oder Anwenders, veranschaulichen und formalisieren. Im Zuge des DIAS Projekts an der Königlichen Bibliothek der Niederlande[26] wurde dieses Konzept vorgestellt. Die Abbildung 15.3.4.3 zeigt einen typischen View-Path ausgehend vom Primärobjekt, über seine Erstellungsapplikation, das durch diese erforderte Betriebssystem und daraus resultierendem Hardwareemulator.

26 siehe dazu das Konzept des Preservation-Managers in http://www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html

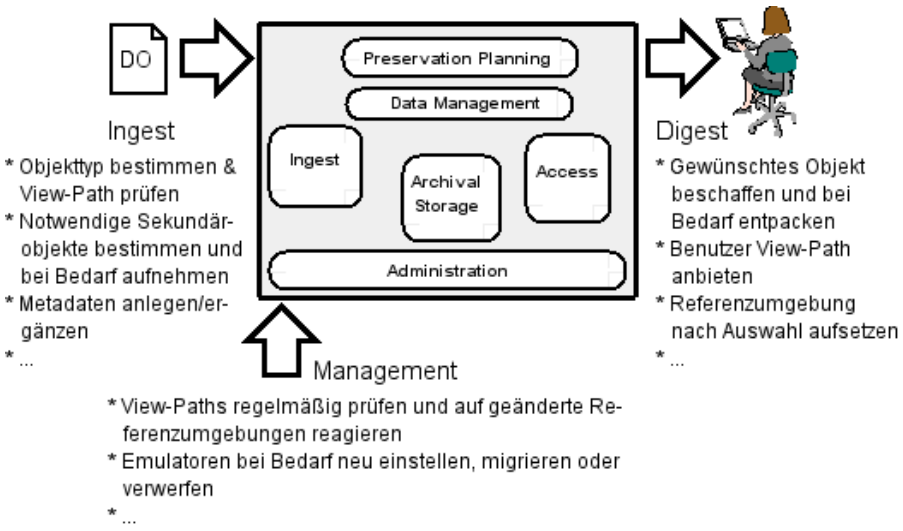


Abbildung 15.3.4.3: Verschiedene Aufgabenstellungen für das Management eines Langzeitarchivs

Für ein digitales Langzeitarchiv erwachsen daraus verschiedene Aufgaben, die von der Bestimmung des Objekttyps beim Einstellen in das Archiv, der Generierung und Ablage der benötigten Metadaten bis zur Überprüfung auf Vorhandensein oder der Herstellbarkeit einer Ablaufumgebung reichen. Abbildung 15.3.4.4 zeigt ein digitales Objekt, das für seine Betrachtung eine Reihe von Sekundärobjekten braucht - angefangen von der Applikation, mit der es erstellt wurde, über das Betriebssystem, das diese ausführen kann. Letzteres wird sich je nach Alter nicht mehr auf moderner Hardware ausführen lassen und läuft deshalb in einem geeigneten Hardwareemulator ab.

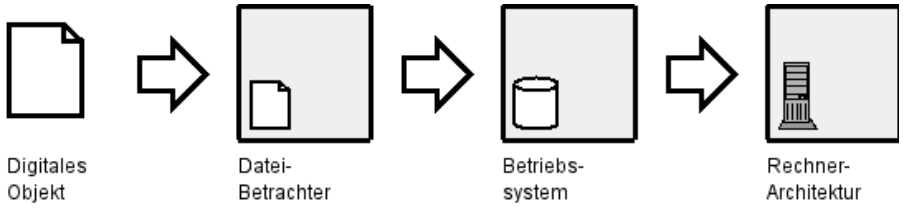


Abbildung 15.3.4.4: Ein typischer View-Path ausgehend von einem digitalen Primärobjekt

Für das OAIS-Management leiten sich daraus verschiedene Aufgaben ab. Eine Rolle spielt beispielsweise die Klassifikation der Objekte nach ihrem Objekttyp (neben ihren Metadaten), da dieser über die später notwendigen Schritte für den Zugriff mitbestimmt. Ebenso können je nach Benutzergruppe oder Ein-

richtung verschiedene Kriterien existieren, nach denen View-Paths bestimmt werden. Es lassen sich drei wesentliche Phasen voneinander unterscheiden:

- Notwendige Handlungen bei der Objektaufnahme ins Archiv (Ingest)
- Regelmäßige Schritte und Arbeitsabläufe während des Archivbetriebs
- Handlungsanweisungen für den Objektzugriff durch den beziehungsweise die Objektausgabe an den Endbenutzer (Digest)

Zur Bestimmung des Objekttyps existieren verschiedene Möglichkeiten; ein über das Internet nutzbarer Dienst mit einer breiten Unterstützung ist beispielsweise PRONOM. Die Auswahl über die in das Softwarearchiv einzustellenden und später zu pflegenden Emulatoren entscheidet wesentlich über die Nachnutzbarkeit der Primärobjekte des Langzeitarchivs. Deshalb ist bei der Einstellung eines digitalen Objekts in das Archiv zu bestimmen, ob für diesen Objekttyp ein View-Path bestimmt werden kann und auf welche später nachzubildende Rechner-Plattform dieser zeigt. Umgekehrt kann das beim Ingest bedeuten, dass ein Objekt zurückgewiesen werden kann oder es nur in Erwartung auf eine später verfügbare Lösung eingestellt wird.

Aus Sicht des Archivmanagements unterscheiden sich Emulatoren nicht wesentlich von den Primärobjekten. Sie werden je nach Archivstrategie und Referenzplattform im Laufe der Zeit obsolet, müssen geeignet migriert oder ihre Nutzungsumgebung erhalten werden. Zu einem gegebenen Zeitpunkt wird ein Primärobjekt von einem Archivbenutzer nachgefragt werden. Dieser erwartet typischerweise eine geeignete Umgebung, in der er das Objekt betrachten oder benutzen kann. Die Benutzergruppen einzelner digitaler Archive und Sammlungen werden sich im Grad ihrer Kenntnisse unterscheiden. Da bei einem durchschnittlichen Nutzer nicht zwingend von einem erfahrenen Computeranwender auszugehen ist, sind Überlegungen zu treffen, wie dieser geeignet an die notwendige Software und ihre Schnittstellen herangeführt werden kann, um mit ihr umgehen zu können. Ebenso kann es notwendig werden eine Reihe von Hilfsmitteln zu präsentieren.

Während der Ausgangspunkt des View-Paths durch das Primärobjekt festgelegt ist, wird sich, erzwungen durch den technologischen Fortschritt und die sukzessive Obsoleszenz vorhandener Rechnerplattformen, der Endpunkt des View-Path im Zeitablauf verschieben. Zudem sind die Längen eines View-Path vom Typ des Primärobjekts abhängig. Generell ergeben sich folgende Szenarien für View-Paths:

- Es gibt zum gegebenen Zeitpunkt einen Weg vom Primärobjekt zu seiner Darstellung oder Ausführung.
- Es existieren mehrere verschiedene View-Paths, diese sind mit geeigneten Metriken zu versehen.

- Es kann Primärobjekte geben, zu denen zu bestimmten Zeitpunkten keine View-Paths vorhanden sind.

Zur sinnvollen Bestimmung der Existenz von View-Paths sollten sie sich deshalb auf bestimmte Referenzumgebungen mit jeweils festgelegten Eigenschaften aus Hard- und Software beziehen. Einen View-Path kann man sich damit als Entscheidungsbaum vorstellen, an dessen Wurzel das Primärobjekt steht. Zur Veranschaulichung des Aufbaus der benötigten Ablaufumgebung läßt sich wiederum ein Schichtenmodell vorstellen.

Viele Primärobjekte lassen sich durch mehr als eine Applikation (Viewer) darstellen. Dabei können die Anzeigeergebnisse in Authentizität, Komplexität oder Qualität differieren. Damit ergibt sich eine Pfadverzweigung und auf der Schicht der Applikation eine Auswahl. Ähnliches trifft auf die Anforderung der Applikation nach einem Betriebssystem zu, so dass in dieser Schicht eine weitere Verzweigung auftreten kann. Die Rekursion setzt sich mit dem Betriebssystem und einer möglichen Auswahl an geeigneten Hardware-Emulatoren fort.

Die Modellierung des View-Path in Schichten erfolgt nicht starr: So reduziert sich beispielsweise bei einem digitalen Primärobjekt in Form eines Programms die Zahl der Schichten. Ähnliches gilt für einfache Plattformen, wie Home-Computer, wo keine Trennung zwischen Betriebssystem und Applikation vorliegt. Darüber hinaus können Schichten wiederum gestapelt sein, wenn es erforderlich wird für einen bestimmten Emulator seinerseits eine geeignete Ablaufumgebung herzustellen (vgl. Abbildung 15.3.4.1).

Eine sinnvolle Erweiterung des etwas unbestimmten Ansatzes im starren DIAS-Modell (eine der ersten Realisierungen des View-Path-Konzepts) könnte in der Gewichtung der einzelnen View-Path-Optionen liegen, die durch eine beschreibende Metrik abgebildet werden könnte. Gerade wenn an einem Knoten mehr als eine Option zur Auswahl steht, wäre es sinnvoll:

- Präferenzen des Benutzers beispielsweise in Form der Auswahl der Applikation, des Betriebssystems oder der Referenzplattform zuzulassen.
- Gewichtungen (anhand bestimmter Metriken) vorzunehmen, ob beispielsweise besonderer Wert auf die Authentizität der Darstellung oder eine besonders einfach Nutzung gelegt wird.
- Vergleiche zwischen verschiedenen Wegen zuzulassen, um die Sicherheit und Qualität der Darstellung der Primärobjekte besser abzusichern.
- Den Aufwand abzuschätzen, der mit den verschiedenen View-Path verbunden ist, um bei Bedarf eine zusätzliche ökonomische Bewertung zu erlauben.

Eine Schlussfolgerung könnten mehrdimensionale Metriken sein, die mit den Objektmetadaten gespeichert und durch das Archivmanagement regelmäßig

aktualisiert werden. Da die Entscheidung über die Qualität einer Darstellung oder Ausführung eines Primärobjekts oft nur vom Anwender getroffen werden kann, sollte man Überlegungen anstellen, wie Benutzerrückmeldungen zur Erstellung der Metriken einfließen könnten.

Geeignete *Referenzumgebungen*, als Bezugspunkte für die Darstellung der Primärobjekte, versuchen in möglichst kompakter und gut bedienbarer Form ein ganzes Spektrum von Ablaufumgebungen zur Verfügung stellen zu können. Dabei sollte die Basisplattform möglichst der jeweils aktuellen Hardware mit jeweils üblichen Betriebssystemen entsprechen. Das verhindert einerseits das Entstehen eines Hardwaremuseums mit hohen Betriebskosten. Andererseits findet sich der Benutzer zumindest für das Basissystem in der gewohnten Umgebung wieder.

Jede historische Rechnerplattform, bis zu den heutigen, weist ihre eigenen Komplexitäten auf, die nicht von jedem durchschnittlichen Computeranwender sinnvoll bewältigt werden können. Zudem kann es nicht die Voraussetzung für den Zugriff auf ein bestimmtes Primärobjekt sein, dass die interessierte Person sich mit der Erstellungsumgebung und ihrer Einrichtung auskennt. Je nach Typus ihrer Benutzer werden deshalb Betreiber digitaler Langzeitarchive nach Lösungen suchen, die es erlauben in geeigneter Weise auf ihre jeweiligen Primärobjekte zuzugreifen. In jedem Fall ist eine gewisse Abstraktionsschicht zu schaffen, über die ein Objektzugriff erfolgen kann. Hierfür sind verschiedene Varianten denkbar, die an unterschiedlichen Schnittstellen ansetzen:

- Referenzplattformen aus einer bestimmten Hardware und Software werden als Endpunkte von View-Paths benötigt. Sie können ortsnah zum Langzeitarchiv, beispielsweise parallel oder als Erweiterung der üblichen Recherchesysteme der jeweiligen Gedächtnisorganisation aufgestellt oder in diese integriert werden.
- Sie ließen sich in einem gewissen Umfang virtualisieren, um sie per Web-Browser oder anderen geeigneten und allgemein genutzten Internet-Technologien entfernt zur Verfügung zu stellen.
- Alternativen bestehen darin definierte virtuelle Maschinen, wie Java, zur Ausführung von Emulatoren und ihren enthaltenen Nutzungsumgebungen zu verwenden (Dioscuri). Hierdurch erreicht man einerseits eine breitere Auswahl von Referenzumgebungen bei einer stabileren Schnittstelle zur virtuellen Maschine. Andererseits verlagert man damit das Problem eine Schicht weiter nach unten und hängt nun von der Weiterentwicklung virtuellen Maschine ab.

Eine Referenzumgebung sollte in der Lage sein, neben der jeweiligen Nutzungsumgebung zusätzlich die notwendigen Hinweise zum Aufsetzen und zur Bedienung bereitzustellen, welche einen geeigneten Zugriff auf die Objektmetadaten

beinhalten. Weitere Kriterien liegen in der Güte der Darstellung der Nutzungsumgebung. Wegen ihrer durch die eingesetzten Emulatoren und Viewer spezielleren Anforderungen ist es für die Betreiber von Langzeitarchiven vielfach sinnvoll eine Referenzplattform selbst zu definieren und bereitzustellen. Diese wird sich je nach Anwendung und Gedächtnisorganisation unterscheiden: Bibliotheken und Archive benötigen in erster Linie Viewer für migrierte statische Objekte und Emulatoren für die nicht-migrierten Archivinhalte. Soweit sinnvoll kann diese Aufgabe auf bereits vorhandenen Recherche- oder Anzeigesystemen der Institution untergebracht werden, um den Benutzern einen leichten Zugriff zu erlauben. Ein Datenaustausch mit der Außenwelt kann benötigt werden, wenn Datenarchäologie angeboten werden oder ein Ausdruck eines Dokuments erfolgen soll.
Technische Museen oder Ausstellungen leben eher von interaktiven Objekten. Die Referenzworkstation ist je nach zu zeigendem Exponat zu bestücken; ein Datenaustausch ist üblicherweise nicht vorgesehen. Für Firmen oder Institutionen kann bereits ein X86-Virtualisierer ausreichen, um der Zugreifbarkeit auf den Archivbestand Genüge zu tun. Die erwarteten Objekte sind eher statischer Natur und wurden typischerweise auf PC's verschiedener Generationen erstellt. Generell muss es sich bei den eingesetzten Referenzworkstations nicht um die jeweils allerneueste Hardwaregeneration handeln. Stattdessen sollte die Technologie angestrebt werden, die einen optimalen Austausch erlaubt und den Anforderungen der jeweiligen Nutzer gerecht wird.
Es muss ein ausreichendes Bedienungswissen vorgehalten werden, welches bei speziellen Nutzergruppen wie digitalen Archivaren, auch für recht alte Nutzungsumgebungen erwartet werden kann. Auf diese Weise lassen sich zudem Versionen der Hardware und Betriebssysteme bei allfälligen Generationswechseln der Referenzworkstations überspringen.

15.4 Web-Harvesting zur Langzeiterhaltung von Internet-Dokumenten

Hans Liegmann

(überarbeitete Fassung eines Vortrags auf der 10. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen" - Planungen, Projekte, Perspektiven – Zum Stand der Archivierung elektronischer Unterlagen - Düsseldorf, 14./15. März 2006)

1 Web-Harvesting als Sammelmethode für Internet-Dokumente

Unter Web-Harvesting versteht man das automatisierte Einsammeln von Internet-Dokumenten zum Zwecke der Archivierung in einem digitalen Archiv. Zentrales Element des Web-Harvesting ist eine Software-Komponente (crawler). Diese sucht ausgehend von einer Liste vorgegebener Web-Adressen (URL seed list) die erreichbaren Dokumente auf und speichert sie in einer definierten Zielumgebung ab.
Beim selektiven zielgerichteten Web-Harvesting (focused crawl) besteht das Ziel darin, möglichst vollständige und konsistente Archivkopien genau derjenigen Websites zu erhalten, deren Adressen in der vorgegebenen Liste enthalten sind.

Beim flächigen Web-Harvesting (broad crawl) wird eine vorgegebene Adressliste lediglich als Einstieg in ein Sammelverfahren verwendet, das weitergehend ist. Flächiges Web-Harvesting hat definierte formale Regeln als Auswahlgrundlage der zu archivierenden Websites. Eine typische Regel kann lauten, dass zu archivierende Dokumente Bestandteil eines bestimmten Internet-Bereiches (domain, z.B. „de") sein müssen, um als archivierungswürdig angesehen zu werden.

Unabhängig vom Komplexitätsgrad möglicher Regelformulierungen ist die Grundlage des Sammelverfahrens die Verfolgung von Hyperlinks: aus den aufgefundenen Dokumenten werden wiederum die in ihnen enthaltenen Web-Adressen extrahiert und auf Regelkonformität geprüft. Die Liste der aufzusuchenden URLs wird dann ggf. dynamisch erweitert.

Derzeit gibt es verschiedene Produkte auf dem Markt, die zur Durchführung von Web-Harvesting geeignet sind. Das Angebot ist vorrangig auf die Bedürfnisse des selektiven Harvesting ausgerichtet. Dazu gibt es kommerzielle, Freeware- und Open-Source-Angebote. Diese genügen überwiegend den Anforde-

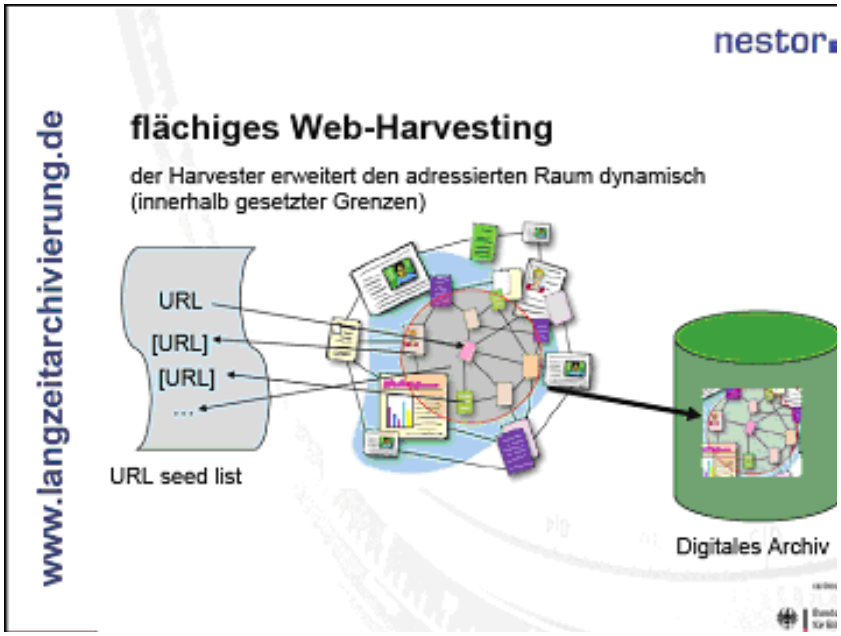


Abbildung 15.4.1

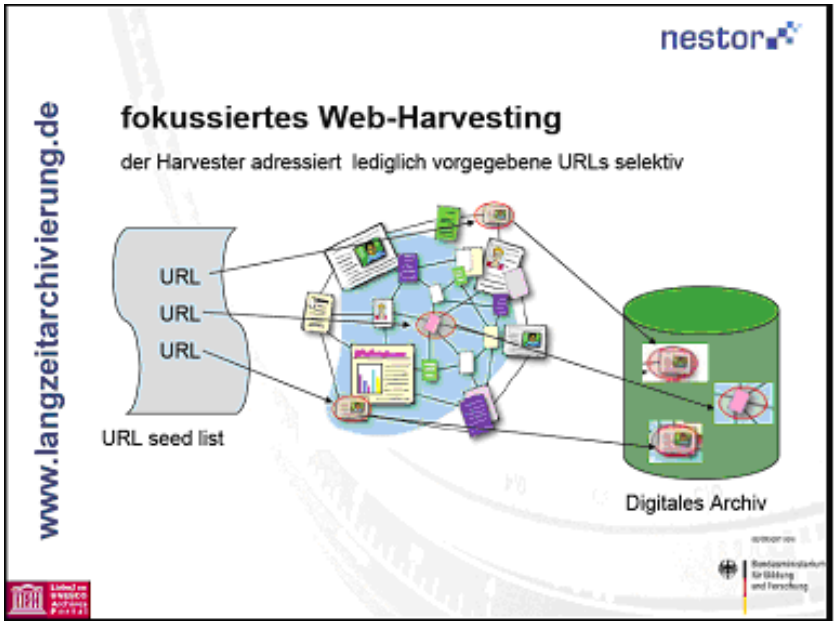


Abbildung 15.4.2

rungen der Langzeitarchivierung nicht, da sie bei der Archivierung der Daten inhaltliche Veränderungen vornehmen.
Flächiges Harvesting unter Berücksichtigung der Authentizität archivierter Objekte wird nur von wenigen Softwareprojekten (z.B. der Crawler HERITRIX des International Internet Preservation Consortium) unterstützt. Bei der Planung produktiver Harvesting-Anwendungen im Massenbetrieb ist zu berücksichtigen, dass kommerzielle Software-Produkte mit garantiertem Leistungsumfang nicht zur Verfügung stehen und ggf. umfangreiche Zusatzinvestitionen notwendig sind, um die gewünschte Funktionalität zu erreichen.
Die aktuelle Anwendungsbreite von Web-Harvesting-Verfahren ist in folgendem Schaubild dargestellt:

Die eingesetzten Verfahren lassen sich in einer Matrix einordnen, die nach den Kriterien „flächig" bis „fokussiert" und „nationale/regionale Auswahl" bis „fachlich/institutionelle" Auswahl aufgebaut ist. Die Aktivitäten von Nationalbibliotheken sind zum Teil flächig angelegt (Sammeln nationaler Adressräume) oder auch durch selektives Vorgehen bestimmt (Auswahl der für einen bestimmten Kulturkreis als relevant bewerteten Internetpräsenzen). Im Bereich der fokussierten Harvesting-Ansätze finden sich fachlich orientierte Beispiele wie z.B. das Projekt DACHS[27], die Vorgehensweise des Deutschen Parlamentsarchivs[28] mit institutioneller Abdeckung und die kooperativen Aktivitäten einiger deutscher Parteienarchive[29].
Bei der Darstellung der Methode soll nicht unerwähnt bleiben, dass die technischen Instrumentarien zur Durchführung zurzeit noch mit einigen Defiziten behaftet sind:

- Inhalte des so genannten „deep web" sind durch Harvester nicht erreichbar. Dies schließt z.B. Informationen ein, die in Datenbanken oder Content Management Systemen gehalten werden. Harvester sind noch nicht in der Lage, auf Daten zuzugreifen, die erst auf spezifische ad-hoc-Anfragen zusammengestellt werden und nicht durch Verknüpfungen statischer Dokumente repräsentiert sind.
- Inhalte, die erst nach einer Authentisierung zugänglich sind, entziehen sich verständlicherweise dem Harvesting-Prozess.

27 http://www.sino.uni-heidelberg.de/dachs [DACHS - Digital Archive for Chinese Studios] (Juni 2006)

28 http://www.bundestag.de/bic/archiv/oeffent/ArchivierungNetzressourcenKlein.pdf [Angela Ullmann; Steven Rösler: Archivierung von Netzressourcen des Deutschen Bundestages] (Juni 2006)

29 http://www.fes.de/archiv/spiegelungsprojekt.htm [Politisches Internet-Archiv] (Juni 2006)

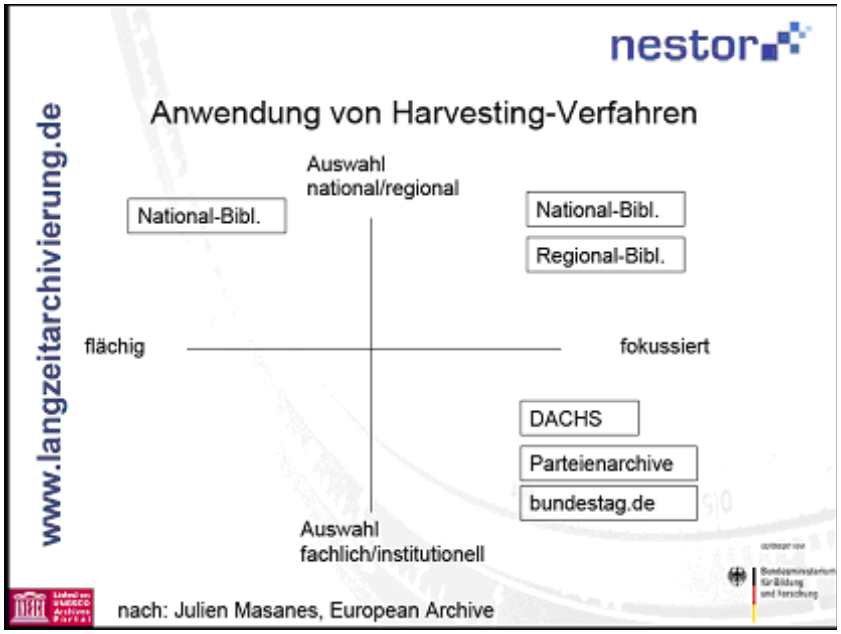


Abbildung 15.4.3

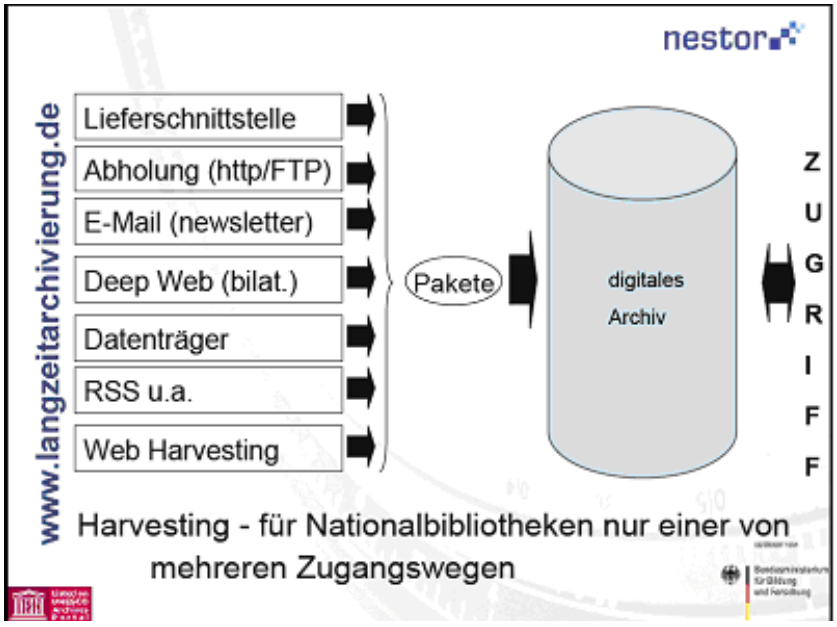


Abbildung 15.4.4

- dynamische Elemente als Teile von Webseiten (z.B. in Script-Sprachen) können Endlosschleifen (crawler traps) verursachen, in denen sich der Harvester verfängt.
- Hyperlinks in Web-Dokumenten können so gut verborgen sein (deep links), dass der Harvester nicht alle Verknüpfungen verfolgen kann und im Ergebnis inkonsistente Dokumente archiviert werden.

Vor allem bei der Ausführung flächigen Web-Harvestings führen die genannten Schwächen häufig zu Unsicherheiten über die Qualität der erzielten Ergebnisse, da eine Qualitätskontrolle aufgrund der erzeugten Datenmengen nur in Form von Stichproben erfolgen kann. Nationalbibliotheken verfolgen deshalb zunehmend Sammelstrategien, die das Web-Harvesting als eine von mehreren Zugangswegen für Online-Publikationen etablieren.

Der individuelle Transfer von Einzeldokumenten über Einlieferschnittstellen oder teilautomatisierte Zugangsprotokolle sowie bilaterale Vereinbarungen mit Produzenten bilden eine wichtige Ergänzung des „vollautomatischen" Sammelverfahrens.

2 Nationalbibliotheken und das World Wide Web

Nationalbibliotheken fassen grundsätzlich alle der im World Wide Web erreichbaren Dokumente als Veröffentlichungen auf und beabsichtigen, ihre Sammelaufträge entsprechend zu erweitern, soweit dies noch nicht geschehen ist. Eine Anzahl von Typologien von Online-Publikationen wurde als Arbeitsgrundlage geschaffen, um Prioritäten bei der Aufgabenbewältigung setzen zu können und der Nutzererwartung mit Transparenz in der Aufgabenwahrnehmung begegnen zu können. So ist z.B. eine Klassenbildung, die mit den Begriffen „druckbildähnlich" und „webspezifisch" operiert, in Deutschland entstanden.[30] In allen Nationalbibliotheken hat die Aufnahme von Online-Publikationen zu einer Diskussion von Sammel-, Erschließungs- und Archivierungsverfahren geführt, da konventionelle Geschäftsgänge der Buch- und Zeitschriftenbearbeitung durch neue Zugangsverfahren, die Masse des zu bearbeitenden Materials und neue Methoden zur Nachnutzung von technischen und beschreibenden Metadaten nicht anwendbar waren. Die neue Aufgabe von Gedächtnisorganisationen, die langfristige Verfügbarkeit digitaler Ressourcen zu gewährleisten, hat zu neuen Formen der Kooperation[31] und Verabredungen zur Arbeitsteilung geführt.

30 http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/digitalebib1104.pdf [Auswahlkriterien für das Sammeln von Netzpublikationen im Rahmen des elektronischen Pflichtexemplars] (Juni 2006)

31 http://www.langzeitarchivierung.de [nestor - Kompetenznetzwerk Langzeitarchivierung]

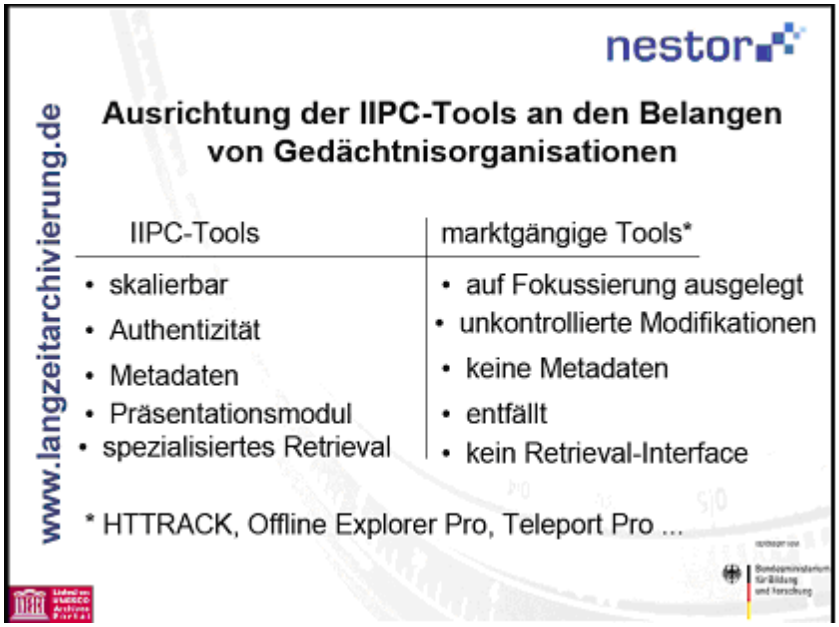


Abbildung 15.4.5

Eine Umfrage der IFLA[32] im Jahr 2005 hat ergeben, dass 16 Nationalbibliotheken Web-Harvesting praktizieren, 11 davon flächiges Harvesting in unterschiedlichen Stadien der Produktivität. 21 Nationalbibliotheken setzen parallel oder ausschließlich andere Verfahren zur Sammlung von Online-Publikationen ein. Die Ergebnisse von Web-Harvesting-Verfahren sind aus urheberrechtlichen Gründen fast ausschließlich nur in den Räumen der jeweiligen Nationalbibliothek zugänglich.

Ein „Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications“[33] der Conference of European National Librarians (CENL) und der Federation of European Publishers (FEP) hat folgende Prinzipien im Umgang zwischen Verlagen und nationalen Archivbibliotheken empfohlen (unabhängig davon, ob sie gesetzlich geregelt werden oder nicht):

(Juni 2006)

32 http://www.ifla.org/ [International Federation of Library Organisations] (Juni 2006)

33 http://www.sne.fr/1_sne/pdf_doc/FINALCENLFEPDraftStatement050822.doc [Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications] (Juni 2006)

- Ablieferung digitaler Verlagspublikationen an die zuständigen Bibliotheken mit nationaler Archivierungsfunktion
- Geltung des Ursprungsland-Prinzip für die Bestimmung der Depotbibliothek, ggf. ergänzt durch den Stellenwert für das kulturelle Erbe einer europäischen Nation
- Einschluss von Publikationen, die kontinuierlich verändert werden (websites) in die Aufbewahrungspflicht
- nicht im Geltungsbereich der Vereinbarung sind: Unterhaltungsprodukte (z.B. Computerspiele) und identische Inhalte in unterschiedlichen Medienformen (z.B. Online-Zeitschriften zusätzlich zur gedruckten Ausgabe.

Das Statement empfiehlt, technische Maßnahmen zum Schutz des Urheberrechts (z.B. Kopierschutzverfahren) vor der Übergabe an die Archivbibliotheken zu deaktivieren, um die Langzeitverfügbarkeit zu gewährleisten.

3 Nationale Strategien von Nationalbibliotheken

Die norwegische Nationalbibliothek[34] gibt in ihren Planungen für das Jahr 2005 an, viermal im Jahr ein Harvesting des vollständigen nationalen Adressraumes (.no) durchführen zu wollen. Darüber hinaus sollen Online-Tageszeitungen täglich und Online-Zeitschriften in der Häufigkeit ihrer Erscheinungsweise eingesammelt werden. Online-Publikationen mit einer Bedeutung für das norwegische kulturelle Erbe, die in anderen top level domains (z.B. .com, .org, .net) erscheinen, werden in Auswahl archiviert. Datenbanken und Netzpublikationen, die im deep web erscheinen und derzeit nicht durch automatische Harvesting-Verfahren erreichbar sind, bleiben vorerst unberücksichtigt.

Die amerikanische Library of Congress (LoC) hat im Jahr 2000 das MINERVA-Projekt[35] eingerichtet und mit Web Harvesting experimentiert. Dabei hat sich die LoC auf den Aufbau thematischer Sammlungen von Websites konzentriert. In Kooperation mit dem Internet Archive[36] wurden so z.B. folgende Sammlungen eingerichtet: Wahlen zum 107. Kongress, Präsidentschaftswahlen, 11. September 2001. Vorgesehen ist die Sammlung und Archivierung von Websites zu den Olympischen Winterspielen 2002, dem Irak-Krieg und weiteren Wahlen auf nationaler Ebene. Die Aktivitäten der amerikanischen Nationalbibliothek bei der Bildung thematischer Sammlungen stehen im Einklang mit der Vorge-

34 http://www.nb.no/english [The National Library of Norway] (Juni 2006)

35 www.loc.gov/minerva [MINERVA - Mapping the Internet Electronic Resources Virtual Archive] (Juni 2006)

36 http://archive.org [Internet Archive] (Juni 2006)

hensweise bei ihren Digitalisierungsvorhaben zum „American Memory“[37].
Die australische Nationalbibliothek[38] war Vorreiter für die Anwendung innovativer technischer Methoden bei der selektiven Sammlung kulturell bedeutender Websites in Australien. Das dortige digitale Archiv PANDORA[39] wird seit 1996 betrieben. In einem kooperativen Verfahren wird es arbeitsteilig zusammen mit den australischen State Libraries aufgebaut. Eingesetzt wird fokussiertes Harvesting unter Verwendung der Standard-Software HTTRACK[40]. Die zusätzlich durchgeführte intensive Qualitätskontrolle der zu archivierenden Inhalte kostet personelle Ressourcen: bislang konnten durch das mit der Aufgabe betraute Personal (ca. 6 Stellen) insgesamt etwa 12.000 Websites mit 22.000 „Schnappschüssen“ aufgenommen und mit Metadaten versehen werden. Da vorab von jedem einzelnen Urheber das Einverständnis zur Archivierung und öffentlichen Bereitstellung eingeholt wird, ist PANDORA eines der wenigen Web-Archive weltweit, die über das WWW offen zugänglich sind.
Die Nationalbibliotheken von Neuseeland und Großbritannien haben im Rahmen Ihrer selektiven Aktivitäten zur Archivierung wichtiger Websites ihres jeweiligen nationalen Adressraumes ein „Web Curator Tool“[41] entwickelt, das als Freeware allen interessierten Anwendern zur Begutachtung und Verfügung steht.

4 Das International Internet Preservation Consortium (IIPC)

Das IIPC[42] wurde 2003 gegründet. Ihm gehören elf Nationalbibliotheken und das Internet Archive an. Die Gründungsidee des IIPC ist es, Wissen und Informationen aus dem Internet für zukünftige Generationen zu archivieren und verfügbar zu machen. Dies soll durch weltweiten Austausch und Kooperation aller Gedächtnisorganisationen erreicht werden, die sich der neuen Aufgabe stellen.
Die Aktivitäten des IIPC sind vielschichtig. Internationale Kooperation auf einem technischen Gebiet erfordert Standardisierung. So hat das IIPC Mitte 2005 einen Standardisierungsvorschlag (Internet Draft) für das „Web Archive File Format (WARC)“ vorgelegt. Eine Standardisierung des Archivierungsformates vereinfacht die Entwicklung nachnutzbarer technischer Instrumentarien

37 http://memory.loc.gov/ammem/index.html [The Library of Congress - American Memory] (Juni 2006)

38 http://www.nla.gov.au [National Library of Australia] (Juni 2006)

39 http://pandora.nla.gov.au [PANDORA - Australia's Web Archive] (Juni 2006)

40 http://www.httrack.com [HTTrack Website Copier - Offline Browser] (Juni 2006)

41 http://webcurator.sourceforge.net

42 http://www.netpreserve.org [International Internet Preservation Consortium] (Juni 2006)

unter den IIPC-Partnern und erlaubt auch den Austausch von Datenbeständen zur redundanten Speicherung aus Sicherheitsgründen.
Unter dem Projektnamen „HERITRIX“[43] arbeiten die IIPC-Partner an einem Web-Harvester, der allen interessierten Anwendern als Open Source Software frei zur Verfügung steht. HERITRIX tritt mit dem Anspruch an, eine skalierbare und ausbaufähige Software zu entwickeln, die (im Gegensatz zu marktüblichen Produkten) Ergebnisse mit Archiv-Qualität liefert. Standard-Produkte erzeugen normalerweise Veränderungen in den lokalen Kopien von Websites, die den Authentizitätsansprüchen von Gedächtnisorganisationen zuwiderlaufen.
Mit NutchWAX[44] (Nutch & Web Archive Extensions) haben IIPC-Partner eine Suchmaschine für den Einsatz in der Web-Archiv-Umgebung vorbereitet. Damit wird es möglich, die Erwartungen von Web-Archiv-Nutzern im Hinblick auf den Suchkomfort durch die Integration von Standard-Suchmaschinentechnologie zu erfüllen.
WERA[45] (Web Archive Access) ist der Prototyp einer Zugriffskomponente, die als Endnutzer-Schnittstelle den Zugang zum digitalen Archiv erlaubt. Im Gegensatz zu marktüblichen Standard-Tools (z.B. HTTRACK) sind die Ergebnisse des Harvesters HERITRIX als Datenpakete im WARC-Format nicht ohne weiteres von Endnutzern zu betrachten. WERA ergänzt die üblichen Suchfunktionen um die Möglichkeit, einen Zeitpunkt für die Auswahl des gewünschten Schnappschusses im Archiv angeben zu können. Damit ist es möglich, mehrere in zeitlicher Abfolge geharvestete Schnappschüsse zusammen zu verwalten und Endnutzern komfortable Suchmöglichkeiten unter Einbeziehung der Zeitachse zu bieten.
Das IIPC sucht auch nach Lösungen, die oben genannten Defizite automatischer Web-Harvesting-Verfahren auszugleichen. Mit „DeepARC“[46] wurde ein grafischer Editor vorgelegt, der es erlaubt, Strukturen aus relationalen Datenbanken in ein XML-Schema abzubilden. Der Transfer wichtiger Inhalte aus dem deep web kann unter Nutzung dieses Tools durch bilaterale Vereinbarungen zwischen Datenbankbetreibern und Archiven geregelt und unterstützt werden.
Zusammenfassend drückt das folgende Schaubild aus, dass die Tools des IIPC explizit an den Belangen von Gedächtnisorganisationen ausgerichtet sind, die an der Langzeitarchivierung von WWW-Inhalten interessiert sind.

5 Ein Blick nach Deutschland

Eine Anzahl von Aktivitäten in Deutschland hat sich der Aufgabe „Langzeiter-

43 http://crawler.archive.org/ [HERITRIX] (Juni 2006)

44 http://archive-access.sourceforge.net/projects/nutch [NutchWAX] (Juni 2006)

45 http://archive-access.sourceforge.net/projects/wera [WERA] (Juni 2006)

46 http://deeparc.sourceforge.net [DeepARC] (Juni 2006)

haltung von Internetressourcen“ angenommen. Die Internetpräsenz des Projekts „nestor - Kompetenznetzwerk Langzeitarchivierung“[47] listet in der Rubrik „Projekte“ folgende Institutionen und Vorhaben auf, die sich im engeren Sinne mit der Sammlung und Archivierung von WWW-Ressourcen befassen: Parlamentsarchiv des Deutschen Bundestages, Baden-Württembergisches Online-Archiv, Digital Archive for Chinese Studies (Heidelberg), edoweb Rheinland-Pfalz, Archiv der Webseiten politischer Parteien in Deutschland und das Webseitenarchiv des Zentralarchivs zur Erforschung der Geschichte der Juden in Deutschland. Nähere Angaben und weiterführende Hinweise sind auf www.langzeitarchivierung.de zu finden.

Die Deutsche Nationalbibliothek hat in den vergangenen Jahren vor allem auf die individuelle Bearbeitung von Netzpublikationen und das damit erreichbare hohe Qualitätsniveau im Hinblick auf Erschließung und Archivierung gesetzt. Eine interaktive Anmeldeschnittstelle kann seit 2001 zur freiwilligen Übermittlung von Netzpublikationen an den Archivserver info-deposit.d-nb.de[48] genutzt werden. Im Herbst 2005 wurde zum Zeitpunkt der Wahlen zum Deutschen Bundestag in Kooperation mit dem European Archive[49] ein Experiment durchgeführt, um Qualitätsaussagen über die Ergebnisse aus fokussiertem Harvesting zu erhalten.

47 http://www.langzeitarchivierung.de [nestor - Kompetenznetzwerk Langzeitarchivierung] (Juni 2006)

48 http://info-deposit.d-nb.de [Archivserver der Deutschen Nationalbibliothek] (Februar 2007)

49 http://europarchive.org [European Archive] (Februar 2007)

15.5 Wissenschaftliche Primärdaten

Jens Klump

Einführung

Der Begriff „Primärdaten" sorgt immer wieder für Diskussion, denn die Definition des Begriffs ist sehr von der eigenen Rolle in der wissenschaftlichen Wertschöpfungskette bestimmt. Für den einen sind „Primärdaten" der Datenstrom aus einem Gerät, z.B. einem Satelliten. In der Fernerkundung werden diese Daten „Level 0" Produkte genannt. Für andere sind „Primärdaten" zur Nachnutzung aufbereitete Daten, ohne weiterführende Prozessierungsschritte. Wieder andere differenzieren nicht nach Grad der Verarbeitung sondern betrachten alle Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren, als Primärdaten.

Welche Definition des Begriffs man auch wählt, wissenschaftliche Primärdaten sind geprägt durch ihre Herkunft aus experimentellem Vorgehen, d.h. anders als Daten aus Arbeitsabläufen der Industrie oder Verwaltung stammen wissenschaftliche Primärdaten überwiegend aus informellen Arbeitabläufen, die immer wieder ad hoc an die untersuchte Fragestellung angepasst werden (Barga und Gannon, 2007). Da in den meisten Fällen keine Formatvorgaben vorhanden sind, werden wissenschaftliche Primärdaten in einer Vielfalt von Dateiformaten hergestellt, die semantisch selten einheitlich strukturiert und nur lückenhaft mit Metadaten beschrieben sind. Diese Faktoren stellen für die digitale Langzeitarchivierung von wissenschaftlichen Primärdaten eine größere Herausforderung dar, als die Datenmenge, auch wenn diese in einzelnen Fällen sehr groß sein kann.

Für den Forscher liegt es nicht im Fokus seines wissenschaftlichen Arbeitens, Daten zu archivieren und zugänglich zu machen, denn bisher bestehen keine Anreize an Wissenschaftler, zumindest Daten, die Grundlage einer Veröffentlichung waren, für andere zugänglich zu machen (Klump et al., 2006). Nur an sehr wenigen Stellen besteht heute im wissenschaftlichen Veröffentlichungssystem oder in der Forschungsförderung die Pflicht, Forschungsdaten für andere zugänglich zu machen. Darüber hinaus ist nicht geklärt, wer für die Langzeitarchivierung von wissenschaftlichen Primärdaten verantwortlich ist und wie diese Leistung finanziert wird (Lyon, 2007). Dies führt zu Defiziten im Management und in der Archivierung wissenschaftlicher Daten mit möglichen negativen Folgen für die Qualität der Forschung (Nature Redaktion, 2006).

Durch eine Reihe von Aufsehen erregenden Wissenschaftsskandalen in den

neunziger Jahren des 20. Jahrhunderts sah sich die Deutsche Forschungsgemeinschaft (DFG) gezwungen, „Empfehlungen für einen gute wissenschaftliche Praxis“ auszusprechen (DFG, 1998), die in vergleichbarer Form auch von anderen Wissenschaftsorganisationen übernommen wurden. In ihren Empfehlungen bezieht sich die DFG auf Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren. Sie verlangt von ihren Zuwendungsempfängern, dass diese Daten für mindestens zehn Jahre auf geeigneten Datenträgern sicher aufbewahrt werden müssen (DFG, 1998, Empfehlung 7). Für die einzelnen Disziplinen ist der Umgang mit Daten im einzelnen zu klären, um eine angemessene Lösung zu finden (DFG, 1998, Empfehlung 1). Diese Policy dient jedoch in erster Linie einer Art Beweissicherung; über Zugang zu den Daten und ihre Nachnutzbarkeit sagen die Empfehlungen nichts aus. Zudem ist bisher noch kein Fall bekannt geworden, in dem die DFG Sanktionen verhängt hätte, allein weil der Pflicht zur Archivierung von Daten nicht nachgekommen wurde.
Trotz der Empfehlungen für eine gute wissenschaftliche Praxis sind kohärente Datenmanagementstrategien, Archivierung von wissenschaftlichen Primärdaten und, soweit möglich, Zugang zu Daten meist nur in größeren Forschungsverbünden zu finden, die für Erfolge in der Forschung auf enge Zusammenarbeit angewiesen sind, oder in Fällen, in denen es gesetzliche Vorgaben für den Umgang mit Daten gibt. Wie schon in der Diskussion um den Offenen Zugang zu wissenschaftlichem Wissen (Open Access) zeigt sich hier, dass eine Policy nur wirksam ist, wenn sie eine Verpflichtung mit sich bringt und gleichzeitig Anreize zur Zusammenarbeit bietet (Bates et al., 2006).
Um das Ziel einer nachhaltigen digitalen Langzeitarchivierung von wissenschaftlichen Primärdaten zu erreichen, muss sowohl eine organisatorische Strategie verfolgt werden, die Langzeitarchivierung von Daten zu einem anerkannten Beitrag zur wissenschaftlichen Kultur macht und die gleichzeitig von einer technischen Strategie unterstützt wird, die den Akteuren für die digitalen Langzeitarchivierung von wissenschaftlichen Primärdaten geeignete Werkzeuge in die Hand gibt. Mit dazu gehören eine Professionalisierung des Datenmanagements und der digitalen Langzeitarchivierung von Forschungsdaten auf Seiten der Projekte und Archive.

Organisatorische Strategien

Auf Grund der enormen Summen, die jährlich für die Erhebung wissenschaftlicher Daten ausgegeben werden, beschäftigt sich die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) bereits seit einigen Jahren mit der Frage, wie mit Daten aus öffentlich geförderter Forschung umgegangen werden sollte. Auf dem Treffen der Forschungsminister im Januar

2004 wurde beschlossen, dass der Zugang zu Daten aus öffentlich geförderter Forschung verbessert werden muss (OECD, 2004). Mit diesem Mandat im Hintergrund befragte die OECD die Wissenschaftsorganisationen ihrer Mitgliedsländer zu deren Umgang mit Forschungsdaten. Aus dem Ergebnissen der Befragung wurde eine Studie verfasst und im Dezember 2006 verabschiedete der Rat der OECD eine „Empfehlung betreffend den Zugang zu Forschungsdaten aus öffentlicher Förderung" (OECD, 2006). Diese Empfehlung ist bindend und muss von den Mitgliedsstaaten der OECD in nationale Gesetzgebung umgesetzt werden, die Umsetzung wird von der OECD beobachtet. In Abschnitt M der Empfehlung wird vorgeschlagen, dass schon bei der Planung von Projekten eine nachhaltige, langfristige Archivierung der Daten berücksichtigt wird.
Parallel dazu, und mit mehr Aufsehen in der Öffentlichkeit, wurde im Oktober 2003 von den Wissenschaftsorganisationen die „Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen" veröffentlicht (Berliner Erklärung, 2003), deren Schwerpunkt auf dem Zugang zu wissenschaftlicher Literatur für Forschung und Lehre liegt. In ihre Definition des offenen Zugangs bezieht die „Berliner Erklärung" auch Daten und Metadaten mit ein. Die Langzeitarchivierung ist hier ein Mittel zum Zweck, das den offenen Zugang zu wissenschaftlichem Wissen über das Internet auf Dauer ermöglichen soll. Aufrufe dieser Art wurden stets begrüßt, aber blieben leider ohne Folgen (Zerhouni, 2006). Dieses Problem betrifft die Institutional Repositories des Open Access genauso wie die Datenarchive. Es sollte daher geprüft werden, inwiefern die Strategien, die bei der Umsetzung von Open Access angewandt werden, sich auch auf den offenen Zugang zu Daten anwenden lassen (Bates et al., 2006; Sale, 2006).
Wenngleich es einige Policies gibt, die den Zugang zu Daten ermöglichen sollen, so hat sich erst recht spät die Erkenntnis durchgesetzt, dass die digitale Langzeitarchivierung von Forschungsdaten eine Grundvoraussetzung des offenen Zugangs ist. Eine umfangreiche Studie wurde dazu bereits in der ersten Förderphase des Projekts nestor erstellt (Severiens und Hilf, 2006). Eine ähnliche Studie wurde auch für das britische Joint Information Systems Committee (JISC) veröffentlicht (Lord und Macdonald, 2003) und das Thema in einer weiteren Studie vertieft (Lyon, 2007). Einzelne Systeme, die als Best-Practice Beispiele gelten dürfen, da sie die Voraussetzungen von Offenem Zugang und vertrauenswürdiger digitaler Langzeitarchivierung erfüllen, existieren bereits.
Die Finanzierung der digitalen Langzeitarchivierung von Forschungsdaten ist eine offene Frage, denn bislang gab es für Datenmanagement jenseits des Projektendes weder die notwendigen finanziellen Mittel, noch waren Antragsteller verpflichtet einen entsprechenden Plan vorzulegen. Hier tritt bei den Förder-

organisationen inzwischen ein Umdenken ein, wenngleich es im aktuellen Regelwerk der Forschungsförderung schwierig ist, Infrastruktur für die digitale Langzeitarchivierung von wissenschaftlichen Primärdaten aufzubauen. Durch die Umsetzung der „Empfehlung betreffend den Zugang zu Forschungsdaten aus öffentlicher Förderung“ (OECD, 2006) kann damit gerechnet werden, dass hier neue Möglichkeiten für den Aufbau von wissenschaftlichen Datenzentren und –archiven entstehen werden.

Technische Strategien

Voraussetzung für die digitale Langzeitarchivierung wissenschaftlicher Primärdaten ist, dass es vertrauenswürdige Archive gibt, die diese Aufgabe übernehmen können. Diese Aufgabe wird bereits in einigen Disziplinen von Datenzentren übernommen und auch die Weltdatenzentren des International Council of Scientific Unions (ICSU WDCs) haben sich dieser Aufgabe verpflichtet. In den vielen Fällen, in denen es kein disziplinspezifisches Datenzentrum und –archiv gibt, fehlen Konzepte für eine digitale Langzeitarchivierung von wissenschaftlichen Primärdaten. Eine mögliche Lösung wäre, in Analogie zur Open Archive Initiative, für diese Daten lokale Institutional Repositories aufzubauen (Lyon, 2007). Die Herausforderungen liegen dabei weniger bei den Archivsystemen, wo sie oft vermutet werden, sondern häufiger im Zusammenspiel der Prozesse des Managements von Forschungsdaten und der digitalen Langzeitarchivierung. So beziehen sich nur wenige Datenarchive in der Organisation ihrer Archivprozesse auf das OAIS-Referenzmodell (OAIS, 2002), das die Prozesse der digitalen Langzeitarchivierung beschreibt (Lyon, 2007).
Besondere Herausforderungen an die digitale Langzeitarchivierung von Forschungsdaten erwachsen aus Grid- und eScience-Projekten, die sich auf den ersten Blick in vielen Aspekten nicht wesentlich von anderen Datenproduzierenden Forschungsprojekten unterscheiden. Die enorm großen Datenmengen, die in Grid-Projekten erzeugt und verarbeitet werden und die hohe Komplexität von Daten aus eScience-Projekten lassen jedoch vermuten, dass aus diesen Projekttypen neuartige Anforderungen an die digitale Langzeitarchivierung erwachsen (Hey und Trefethen, 2003). Gerade wegen dieser extremen Anforderungen an Prozessierungs- und Speicherressourcen und zusätzlichen Managementvorkehrungen durch Virtualisierung der Ressourcen sind Communities, die große Datenmengen erzeugen oder verarbeiten, in der Anwendung von Grid-Technologien vergleichsweise weit fortgeschritten. Astrophysik, Klimaforschumg, biomedizinische Forschung, und andere Communities mit rechenintensiven Verfahren der Datenverarbeitung wenden bereits seit einiger Zeit Grid-Technologien an.

Die enorm großen Datenmengen erfordern von den Grid-Projekten konsistente Richtlinien für die Auswahl der Daten, die für lange Zeiträume archiviert werden sollen. Ähnlich wie in den Richtlinien des British Atmospheric Data Centre (Lyon, 2007) wird in den Projekten evaluiert, ob die Daten grundsätzlich und mit vertretbarem Aufwand neu generiert werden können, und ob die Daten in der vorliegenden Form nachnutzbar sind (Klump, in prep.).

Langzeitarchive für wissenschaftliche Primärdaten und organisatorische Rahmenbedingungen in den Instituten und bei der Forschungsförderung sind notwendige Voraussetzungen für die digitale Langzeitarchivierung von wissenschaftlichen Primärdaten. Sie müssen aber auch durch technische Lösungen unterstützt werden, die die Mitwirkung durch die Wissenschaftler an der digitalen Langzeitarchivierung von wissenschaftlichen Primärdaten so einfach wie möglich gestalten, so dass sie sich möglichst nahtlos in die wissenschaftlichen Arbeitsabläufe einfügt. Ein Beispiel dafür ist die Beschreibung der Forschungsdaten durch Metadaten. Erstellen und Pflege von Metadaten stellt eine enorme Hürde dar, denn die notwendigen Metadatenprofile sind meist komplex, sie manuell zu erstellen ist aufwendig (Robertson, 2006). In der Praxis hat sich gezeigt, dass das Management von Daten und Metadaten eine bessere Chance zum Erfolg hat, wenn das Erstellen und Pflegen von Metadaten weitgehend automatisiert ist. Ein hoher Grad an Technisierung des Datenmanagements erlaubt den Wissenschaftlern, sich ihrem eigentlichen Tätigkeitsschwerpunkt, der Forschung, zu widmen. In den vom Bundesministerium für Bildung und Forschung geförderten Projekten C3-Grid (Kindermann et al., 2006) und TextGrid sind sowohl für die Naturwissenschaften, als auch für die Geisteswissenschaften vorbildliche Verfahren für die Erzeugung und Verwaltung von Metadaten entwickelt worden.

Während bereits die inhaltliche Beschreibung der zu archivierenden Daten durch Metadaten eine Hürde darstellt, kommen für eine spätere Nachnutzung weitere Probleme hinzu. Vielfach trifft man auf das Missverständnis, dass die Angabe des MIME-Type eine ausreichende Beschreibung des Dateiformats und seiner Nutzung sei. Ein Archivsystem müsste jedoch nicht nur den MIME-Type der archivierten Daten kennen, sondern auch deren semantische Struktur und ihr technisches Format. Die semantische Struktur maschinenlesbar zu dokumentieren ist eine Grundvoraussetzung für die in Zukunft geforderte Interoperabilität der Archivsysteme (Klump, in prep.). Zusätzlich müssen sich die Archivbetreiber und ihre Nutzer darüber verständigen, welche Dateiformate archiviert werden, denn nicht jedes bei den Nutzern populäre Format ist für eine verlustfreie Langzeitarchivierung geeignet (Lormant et al., 2005).

Ungeachtet des in der „Berliner Erklärung“ durch die Universitäten, Wissen-

schafts- und Forschungsförderungsorganisationen geleisteten Bekenntnisses zum Offen Zugang gibt es Gründe, warum manche Daten nicht offen zugänglich sein können. Aus diesem Grund sind Zugriffsbeschränkungen in der digitalen Langzeitarchivierung von wissenschaftlichen Primärdaten ein wichtiges Thema. Die Zugriffsbeschränkungen dienen hierbei nicht primär der Sicherung von Verwertungsrechten, sondern sie sind entweder gesetzlich vorgeschrieben (Datenschutz) oder dienen dem Schutz von Personen oder Objekten, die durch eine Veröffentlichung der Daten Gefährdungen ausgesetzt würden. Für geschützte Datenobjekte müssen Verfahren und Policies entwickelt werden, die auch über lange Zeiträume hinweg zuverlässig die Zugriffsrechte regeln und schützen können (Choi et al., 2006; Simmel, 2004). Auch der Umgang mit „verwaisten" Datenbeständen muss geregelt werden.
Zum Schutz der intellektuellen Leistung der Wissenschaftler sollten Daten in wissenschaftlichen Langzeitarchiven mit Lizenzen versehen sein, die die Bedingungen einer Nachnutzung regeln, ohne dadurch den wissenschaftlichen Erkenntnisgewinn zu behindern. Entsprechende Vorarbeiten sind bereits in den Projekten Creative Commons (CC) und Science Commons (SC) geleistet worden. Zusätzlich zur erwiesenen Praxistauglichkeit können die hier entwickelten Lizenzen auch maschinenlesbar hinterlegt werden, was eine künftige Nachnutzung deutlich vereinfacht. Die Diskussion, welche Lizenzen für Daten empfohlen werden sollten, ist noch offen (Uhlir und Schröder, 2007).

Nachnutzung von Daten

Keine der Infrastrukturen für eine digitale Langzeitarchivierung lässt sich dauerhaft betreiben, wenn es keine Nutzer gibt, denn erst wenn eine Nachfrage der Wissenschaft nach einer digitalen Langzeitarchivierung besteht, können dauerhafte Strukturen entstehen. Im heutigen Wissenschaftsbetrieb sind der Gewinn an Distinktion und Reputation wichtige Motivationskräfte. Digitale Langzeitarchivierung muss als Praxis in der Wissenschaft verankert sein und im selbst verstandenen Eigeninteresse der Wissenschaftler liegen. Die wissenschaftliche Publikation ist dabei ein entscheidendes Medium. Ein möglicher Anreiz, Daten zu veröffentlichen und dauerhaft zugänglich zu machen, ist es daher, die Veröffentlichung von Daten zu formalisieren und als Bestandteil des wissenschaftlichen Arbeitens zu institutionalisieren. Dazu ist nötig, dass die veröffentlichten Daten findbar, eindeutig referenzierbar und auf lange Zeit zugänglich sind. Allerdings werden Datenveröffentlichungen nur dann auch nachgenutzt und zitiert, wenn ihre Existenz den potenziellen Nutzern auch bekannt ist. Ein geeigneter Weg, Daten recherchierbar und zugänglich zu machen, ist ihre Integration in Fachportale und Bibliothekskataloge. Eine entscheidende Voraus-

setzung für die Zitierbarkeit von Daten ist, dass sie eindeutig und langfristig referenzierbar sind.
Da in der Praxis URLs nur kurzlebig sind, werden sie nicht als zuverlässige Referenzen angesehen. Persistente, global auflösbare Identifier, wie z.B. Digital Object Identifier (DOI) oder Universal Resource Names (URN) schließen diese Lücke (Hilse und Kothe, 2006; Klump et al., 2006). In der bisherigen Praxis fehlten bisher wichtige Bestandteile, die eine nachhaltige Publikation von Daten möglich machen. Diese Defizite wurden im DFG-Projekt „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten" (STD-DOI) analysiert. Mit der Einführung von persistenten Identifikatoren für wissenschaftliche Primärdatensätze wurden die Voraussetzungen für eine Publikation und Zitierbarkeit wissenschaftlicher Primärdaten geschaffen (Brase, 2004).

Zusammenfassung

In der Einleitung zum OAIS-Referenzmodell (OAIS, 2002) zur Langzeitarchivierung digitaler Objekte ist treffend formuliert worden, dass ein Archivsystem für digitale Objekte mehr ist, als nur ein technisches System. Das OAIS-Referenzmodell beschreibt es als das Zusammenwirken von Menschen und Systemen mit dem Ziel der Langzeiterhaltung von digitalen Objekten für eine definierte Nutzergruppe. Die digitale Langzeitarchivierung wissenschaftlicher Primärdaten ist daher nicht allein eine technische Herausforderung sondern muss in einen entsprechenden organisatorischen Rahmen eingebettet sein, der im Dialog mit der Wissenschaft gestaltet wird. Der wissenschaftliche Wert, Forschungsdaten für lange Zeit zu archivieren und zugänglich zu machen, ist erkannt worden. In dem Maße, wie die Auswertung von Daten für die Forschung an Bedeutung zunimmt, wird sich auch der Umgang mit Daten in der Forschungspraxis und in der Langzeitarchivierung verändern.

Literatur

Barga, R. und Gannon, D.B., 2007. Scientific versus business workflows. In: I.J. Taylor, E. Deelman, D.B. Gannon und M. Shields (Hrsg.), Workflows for e-Science. Springer-Verlag, London, Großbritannien, S. 9-16.

Bates, M., Loddington, S., Manuel, S. und Oppenheim, C., 2006. Rights and Rewards Project - Academic Survey Final Report, JISC. http://rightsandrewards.lboro.ac.uk/files/resourcesmodule/@random43cbae8b0d0ad/1137423150_SurveyReport.pdf

Berliner Erklärung, 2003. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Berlin.

Brase, J., 2004. Using Digital Library Techniques - Registration of Scientific Primary Data. In: M. Jones, E.A. Fox und R. Shen (Hrsg.), Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Springer-Verlag, Heidelberg, Germany, S. 488-494.

Choi, H.-C. et al., 2006. Trust Models for Community Aware Identity Management, WWW2006, Edinburgh, Großbritannien.

DFG, 1998. Regeln guter wissenschaftlicher Praxis, Deutsche Forschungsgemeinschaft. http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf

Hey, T. und Trefethen, A., 2003. The data deluge: an eScience perspective. In: F. Berman, T. Hey und G.C. Fox (Hrsg.), Grid Computing - Making the Global Infrastructure Reality. Wiley & Sons, Ltd., New York, NY, USA, S. 409-435.

Hilse, H.-W. und Kothe, J., 2006. Implementing Persistent Identifiers, Consortium of European Research Libraries, London, Großbritannien. http://www.knaw.nl/ecpa/publ/pdf/2732.pdf

Kindermann, S., Stockhause, M. und Ronneberger, K., 2006. Intelligent Data Networking for the Earth System Science Community. In: W. Bühler (Hrsg.), German eScience Conference. Max Planck Digital Library, Baden-Baden.

Klump, J., in prep. Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten. nestor-Materialien, Kompetenznetzwerk Langzeitarchivierung (nestor), Göttingen.

Klump, J. et al., 2006. Data publication in the Open Access Initiative. Data Science Journal, 5: 79-83. doi:10.2481/dsj.5.79

Lord, P. und Macdonald, A., 2003. e-Science Curation Report - Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, JISC. http://www.jisc.ac.uk/uploaded_documents/escienceReportFinal.pdf

Lormant, N., Huc, C., Boucon, D. und Miquel, C., 2005. How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?, Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005), Edinburgh, Großbritannien, S. 11.

Lyon, L., 2007. Dealing with Data: Roles, Rights, Responsibilities and Relationships, UKOLN, Bath, Großbritannien. http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf

Nature Redaktion, 2006. A fair share. Nature, 444(7120): 653-654. doi:10.1038/444653b

OAIS, 2002. Reference Model for an Open Archival Information System (OAIS). Blue Book., CCSDS 650.0-B-1, Consultative Committee for Space Data Systems, Greenbelt, MD, USA. http://public.ccsds.org/publications/archive/650x0b1.pdf

OECD, 2004. Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué, Organisation for Economic Co-operation and Development, Paris, Frankreich. http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

OECD, 2006. Recommendation of the Council concerning Access to Research Data from Public Funding, C(2006)184, Organisation for Economic Co-operation and Development, Paris, Frankreich. http://webdomino1.oecd.org/horizontal/oecdacts.nsf/Display/3A5FB1397B5ADFB7C12572980053C9D3?OpenDocument

Robertson, R.J., 2006. Evaluation of metadata workflows for the Glasgow ePrints and DSpace services, University of Strathclyde, Glasgow, Großbritannien. http://hdl.handle.net/1905/615

Sale, A., 2006. The acquisition of Open Access research articles. First Monday, 11(10). http://firstmonday.org/issues/issue11_10/sale/index.html

Severiens, T. und Hilf, E.R., 2006. Zur Entwicklung eines Beschreibungsprofils für eine nationale Langzeit-Archivierungs-Strategie - ein Beitrag aus der Sicht der Wissenschaften. nestor Materialien, 7, nestor - Kompetenznetzwerk Langzeitarchivierung, Göttingen. urn:nbn:de:0008-20051114018

Simmel, D., 2004. TeraGrid Certificate Management and Authorization Policy, Pittsburgh Supercomputing Center, Carnegie Mellon University, University of Pittsburgh, Pittsburg, PA, USA. http://www.teragrid.org/policy/TG-CertPolicy-TG-5.pdf

Uhlir, P.F. und Schröder, P., 2007. Open Data for Global Science. Data Science Journal, 6(Open Data Issue): OD36-53. doi:10.2481/dsj.6.OD36

Zerhouni, E.A., 2006. Report on the NIH public access policy, National Institute of Health, Bethesda, MD, USA. http://publicaccess.nih.gov/Final_Report_20060201.pdf

15.6 Computerspiele

Karsten Huth

Das Computerspiel ist, neben den frühen Datenbanken, eines der ältesten digitalen Artefakte, dass von seiner Natur her als „born digital" zu betrachten ist. Sieht man von dem ersten Vorläufer des Videospiels, einem Ausstellungsstück auf einem „Tag der offenen Tür" der Atomforschung, und dem ersten Wohnzimmergerät ab, beide Geräte beruhten noch auf analoger Technik, so sind alle Video- und Computerspiele technisch betrachtet Computerprogramme. Das IBM Dictionary of Computing ordnet sie der „application software", also der Anwendungssoftware" zu, zu der auch Textverarbeitungsprogramme, Tabellenkalkulationen und andere Office-Programme gezählt werden. Computerspiele bilden dennoch eine Sondergruppe innerhalb der Anwendungssoftware. Mit ihnen wird kein Problem gelöst oder die täglich anfallende Büroarbeit bewältigt. Computerspiele dienen einzig der Unterhaltung und dem Vergnügen des Nutzers. Ihre unterhaltende Funktion hat technische Konsequenzen. Computerspiele müssen sich auf einem wachsenden Markt behaupten und die Aufmerksamkeit der Käufer erregen. Sie operieren deshalb oft am oberen technischen Limit der jeweiligen aktuellen Hardwaregeneration. Überlieferte Beispiele aus den siebziger oder achtziger Jahren mögen gegen die Leistungsfähigkeiten eines aktuellen PCs rührend anmuten, für den Nutzer vergangener Tage waren sie ein Beispiel für rasenden technischen Fortschritt, das nicht selten Begeisterung auslöste. Diese Begeisterung machte den Einzug des Computers in den privaten Haushalt möglich. Sie legte einen Grundstein für unseren alltäglichen Umgang mit der digitalen Medienwelt.

Video- und Computerspiele werden häufig nach ihren Hardware/Software Plattformen klassifiziert. Man unterscheidet:

- die Arcade-Spiele: Automaten, die in Spielhallen stehen und gegen den Einwurf von Geld benutzt werden können. Die Software befindet sich meistens auf austauschbaren Platinen im sogenannten Jamma-Standard.
- die Computerspiele: Spiele die auf Computern gespielt werden, welche nicht ausschließlich zum Spielen gedacht sind. Ein aktuelles Beispiel sind die PCs. In den achtziger Jahren waren die Homecomputer sehr populär. Das früheste Beispiel ist das Spiel „Spacewar" aus dem Jahr 1962, geschrieben für den ersten Minicomputer der Welt, den PDP-1. Die Datenträger für Computerspiele reichen von üblichen Musikkassetten über die ersten Floppydisks bis hin zu den heute gebräuchlichen DVDs. Die Darstellung des Spiels erfolgte damals über den Fernseher, heute über

den PC-Monitor.

- die Videospiele: Plattform ist hierbei die sogenannte „Konsole". Die Konsole ist ein Computer, der einzig zum Spielen dient. Seine Hardware ist deshalb für eine gute grafische Darstellung und eine gute Audio-Ausgabe optimiert. Die Datenträger sind ebenso wie die Software an einen bestimmten Konsolentyp gebunden.
- die tragbaren Videospiele: Die sogenannten Handhelds vereinigen den Computer, den Monitor und das Steuerungsgerät in einem kompakten Taschenformat. Neu hinzugekommen sind die Spiele für Mobiltelefone. Bei manchen Geräten sind die Spiele fest implementiert, bei anderen sind sie über spezielle Datenträger austauschbar.

(vgl. Fritz, J. 1997, s. 81)

Folgende Gründe sprechen für eine nachhaltig betriebene Langzeitarchivierung von Computerspielen:

Wissenschaftliche Forschung: Computer- und Videospiele sind zum interdisziplinären Untersuchungsgegenstand für die Wissenschaft geworden, vor allem in den Bereichen der Pädagogik, Psychologie, Kultur- und Medienwissenschaften. Das „Handbuch Medien Computerspiele" herausgegeben von der Bundeszentrale für politische Bildung verzeichnet im Anhang ca. 400 Titel zum Thema Computerspiele. Diese Zahl der größtenteils deutschen Titel aus dem Jahr 1997 zeigt, dass die wissenschaftliche Untersuchung von Computerspielen keine Randerscheinung ist. Die Artikel des Handbuchs beziehen sich oft auf konkrete Spielsoftware. Während das Zitieren der Literatur in diesen Artikeln nach wissenschaftlichen Regeln abläuft, werden Angaben zu den verwendeten Spielen oft gar nicht oder nur in unzureichender Weise gemacht. Man kann somit die wissenschaftlichen Hypothesen eines Artikels, der spezielle Computerspiele als Gegenstand behandelt, nicht überprüfen. Neben dem Problem des wissenschaftlichen Zitierens besteht natürlich auch das Problem des gesicherten legalen Zugriffs auf ein zitiertes Computerspiel. Streng genommen, ist ohne eine vertrauenswürdige Langzeitsicherung von Computerspielen die Wissenschaftlichkeit der Forschung in diesem Bereich gefährdet.

Kulturelle Aspekte: Die Anfänge des Computerspiels reichen zurück bis in das Jahr 1958. Seitdem hat sich das Computerspiel als eigenständiges Medium etabliert. Zum ersten Mal in der Geschichte könnten wir die Entwicklung einer Medienform, von den ersten zaghaften Versuchen bis zur heutigen Zeit, beinahe lückenlos erhalten und damit erforschen. Es wird allgemein bedauert, dass aus der frühen Stummfilmzeit nur noch ca. 10% des einst verfügbaren Materials erhalten geblieben sind. Der Bestand an Computerspielen wäre noch zu einem

ökonomisch vertretbaren Preis zu erhalten und könnte auch der übrigen Medienforschung dienen.
Als Zeugnis der technischen Entwicklung: Wie bereits erwähnt, testen Computerspiele, wie keine zweite Software, die technischen Fähigkeiten der jeweiligen Hardwaregeneration aus. Sie eignen sich dadurch für eine anschauliche Demonstration des Mooreschen Gesetzes. Zudem wurde bei alter Software Programmiertechniken verwendet, die auf einen sparsamen und ökonomischen Einsatz von Hardware-Ressourcen (Speicherplatz und Rechenzeit) ausgerichtet waren. Diese Techniken wurden im Zuge der Hardwareverbesserungen aufgegeben und vergessen. Niemand kann jedoch sagen, ob sie nicht irgendwann einmal wieder von Nutzen sein werden. <Dooijes>

Die Integration von Video- und Computerspielen in die Medienarchive, Bibliotheken und Museen steht noch aus. Die Erhaltung der frühen Spiele ist der Verdienst von privaten Sammlern und Initiativen, die sich über das Internet gefunden und gebildet haben. Beinahe jede obsolete Spielplattform hat ihre Gemeinde, die mit großem technischen Expertentum die notwendigen Grundlagen für eine langfristige Archivierung schafft. Den wichtigsten Beitrag schaffen die Autoren von Emulatoren, die oft zur freien Verfügung ins Netz gestellt werden. Aber auch das Sammeln von Metadaten, welches oft in umfangreiche Softwareverzeichnisse mündet, die aufwendige Migration der Spielsoftware von ihren angestammten Datenträgern auf moderne PCs sowie das Sammeln des Verpackungsdesigns und der Gebrauchsanleitungen sind notwendige Arbeiten, die unentgeltlich von den Sammlern erbracht werden. Leider bewegen sich die privaten Initiativen oft in einer rechtlichen Grauzone. Die Software unterliegt dem Urheberrecht. Ihre Verbreitung über das Internet, auch ohne kommerzielles Interesse, stellt einen Rechtsbruch dar, selbst wenn die betroffenen Produktionsfirmen schon längst nicht mehr existieren. Besonders die Autoren von Emulatoren werden von der Industrie in eine Ecke mit den aus Eigennutz handelnden Softwarepiraten gestellt. Es soll hier nicht verschwiegen werden, dass es auch Emulatoren gibt, die aktuelle Spielplattformen emulieren und dadurch die Softwarepiraterie fördern. Die Motivation dieser Autoren ist deutlich anders gelagert. Die Sammler von historischen Systemen nutzen die Emulation zur Erhaltung ihrer Sammlungen. Die obsoleten Systeme sind im Handel in dieser Form nicht mehr erhältlich. Zudem hat die Industrie bislang kaum Interesse an der Bewahrung ihrer eigenen Historie gezeigt. Zumindest gibt es innerhalb der International Game Developers Association (IGDA) eine Special Interest Group (SIG), die sich mit dem Problem der digitalen Langzeitarchivierung befassen will.

Beispiele für die Langzeitarchivierung von Computerspielen in den klassischen Institutionen sind rar. Die Universitätsbibliothek in Stanford besitzt wohl die größte Sammlung innerhalb einer Bibliothek. Die Sammlung trägt den Namen des verstorbenen Besitzers: Stephen M. Cabrinety. Sie besteht aus kommerziellen Videospielen, sowie den Originalverpackungen, Gebrauchsanleitungen, gedruckten Materialien und dokumentiert somit einen großen Teil der Geschichte der Computerspiele in der Zeitspanne von 1970-1995. Neben den 6.300 Programmen verfügt die Sammlung über 400 original Hardwareobjekte von Motherboards, Monitoren bis hin zu CPUs. Die Sammlung wird verwaltet von Henry Lowood und ist Teil des "Department. of Special Collections" der Stanford University Library (Lowood, 2004).
Das Computerspielemuseum in Berlin wurde im Februar 1997 eröffnet. Getragen wird das Museum vom Förderverein für Jugend- und Sozialarbeit e.V. Das Museum besitzt rund 8.000 Spiele und ist auf der Suche nach einem neuen Ort für eine permanente Ausstellung seiner Exponate. Zur Zeit (2006/2007) ist das Museum mit der Ausstellung Pong-Mythos in Stuttgart, Leipzig und Bern auf Tournee.
Der Verein „Digital Game Archive" hat sich den Aufbau eines legalen Medienarchivs für Computerspiele zum Ziel gesetzt. Der Nutzer kann die archivierten Spiele über die Internetseite des Archivs beziehen. Alle angebotenen Spiele wurden von den Rechteinhabern zur allgemeinen Verwendung freigegeben. Neben der Erhaltung der Software sammelt das Digital Game Archive auch Informationen zum Thema Computerspielarchivierung und versucht die Geschichte des Computerspiels zu dokumentieren. Die Mitglieder sind Fachleute aus verschiedenen wissenschaftlichen Disziplinen. Sie vertreten den Verein auch auf Fachkonferenzen. Das Digital Game Archive arbeitet eng mit dem Computerspielemuseum Berlin zusammen.
Das Internet Archive hat eine kleine Sektion, die sich der Sammlung von historischen Computerspielen widmet. Diese hat das Classic Software Preservation Project im Januar 2004 ins Leben gerufen. Ziel des Projekts ist die Migration gefährdeter Software von ihren originalen Datenträgern auf aktuelle, nicht obsolete Medien. Nach der Migration werden die Programme solange unter Verschluss gehalten, bis die Rechtslage eine legale Vermittlung der Inhalte erlaubt. Um dieses Vorhaben rechtlich möglich zu machen, erwirkte das Internet Archive eine Ausnahmeregelung vor dem Digital Millenium Copyright Act. Das Copyright Office entsprach den Vorschlägen des Internet Archives und erlaubte die Umgehung eines Kopierschutzes sowie die Migration von obsoleter Software auf aktuelle Datenträger zum Zwecke der Archivierung in Gedächt-

nisorganisationen. Diese Ausnahmeregelung wird 2006 erneut vom Copyright Office geprüft werden.

Es gibt zwei mögliche digitale Erhaltungsstrategien für die Langzeitarchivierung von Computerspielen, wenn man sich zum Ziel gesetzt hat, die Spielbarkeit der Programme zu erhalten. Die Möglichkeit, das Spiel nur durch Bilder (Screenshots) und eine ausreichende Spielbeschreibung zu dokumentieren und einzig diese Dokumentation zu bewahren, soll hier nicht weiter betrachtet werden. Langzeitarchivierung eines Computerspiels in diesem Kapitel heißt: „Der originale Bitstream des Computerspiels muss erhalten bleiben und das Programm soll auch in Zukunft noch lauffähig und benutzbar sein."
Diese Vorgabe schränkt die möglichen Erhaltungsstrategien von vornherein ein. Migration scheidet als langfristige Strategie aus, da sie bei einer Anpassung an eine neue Softwareplattform den Bitstream des Programms verändert. Solche Portierungen von Programmen auf neue Plattformen sind sehr viel aufwendiger als die vergleichbare Konvertierung von Dateien in ein anderes Dateiformat. Bei einer Dateikonvertierung kann ein einzelnes Konverterprogramm unbegrenzt viele Dateien bearbeiten. Bei einer Software-Portierung muss jedes einzelne Programm von Hand umgeschrieben und angepasst werden. Zudem bräuchte man ein hohes technisches Wissen über die obsoleten Programmiersprachen, welches oft nicht mehr verfügbar ist. Die Kosten und der Aufwand für eine langfristige Migrationsstrategie wären somit immens hoch.
Praktiziert werden zurzeit zwei Erhaltungsstrategien. Zum einen die der Hardware Preservation (Computermuseum) und die der Emulation. Beide Strategien erhalten den originalen Bitstream eines Programms. Diese Zweigleisigkeit findet man sowohl in privaten Sammlerkreisen, als auch bei den Computerspiel bewahrenden Institutionen wieder. Befürworter der Hardware Preservation Strategie bemängeln den Verlust des sogenannten „Look and Feel" bei der Emulation. Diese Kritik ist nicht ganz unberechtigt. Ältere Spiele der 8-Bit Hardwaregeneration wurden beispielsweise für die Ausgabe auf einem NTSC oder PAL Fernsehbildschirm konzipiert. Die Betrachtung mittels eines Emulators über einen PC-Monitor gibt nicht zu einhundert Prozent den ursprünglichen Eindruck wieder. Die Farben wirken, je nach Einstellung, auf jedem PC etwas anders. Teilweise ist die Emulation auch nicht vollständig, sodass z.B. die Tonwiedergabe nicht bei allen Sound-Effekten glückt. Manche Emulatoren bieten zusätzlich eine Anpassung des Bildes an die alten NTSC- oder PAL-Verhältnisse, um Abweichung des „Look and Feel" zu kompensieren. Jenseits von Bild und Ton bleibt aber noch das Problem der Steuerung. Die originalen Steuerungsgeräte (Joystick, Paddle usw.) werden bei einer Emulation auf dem

PC durch die dort vorhandenen Steuerungsgeräte Tastatur und Maus ersetzt. Dies kann zu einem abweichenden Spielerlebnis und Ergebnis führen. Manche Spiele sind mit PC-Tastatur oder Maus nur sehr schwer oder auch gar nicht zu Bedienen. Wir werden später beim Thema „notwendige Metadaten" näher auf dieses Problem eingehen.
Bei der Hardware Preservation muss man sich hingegen fragen, ob es sich hierbei überhaupt um eine Langzeitarchivierungsstrategie handelt. Es dürfte auf lange Sicht hin unmöglich sein, die originale Hardware und die dazugehörigen Datenträger lauffähig zu halten. Einige Datenträger, z.B. EPROMS haben sich als sehr haltbar erwiesen, andere Datenträger z.B. Floppy-Disks halten bestenfalls 10 Jahre. Regelmäßiges überspielen der Programme auf frische Datenträger des gleichen Typs als Strategie zur Lebensverlängerung scheidet aus, da die betreffenden Datenträgertypen obsolet geworden sind und somit nicht mehr über den Handel nachproduziert werden. Somit bleibt nur die Emulation als erfolgversprechende Langzeitstrategie.

Zur Zeit gibt es noch kein funktionierendes Langzeitarchiv für Computerspiele, das den kompletten Anforderungen des OAIS Funktionsmodels entspricht. Im folgenden Abschnitt wird in einfachen Schritten ein OAIS-konformes Modell für ein Computerspielarchiv entworfen. Die einzelnen Abschnitte sind dementsprechend in Ingest (Accession/Erfassung), Data Management/Archival Storage (Erschließung/Magazin), Access (Benutzung) unterteilt. Wenn möglich, werden zu den einzelnen Abschnitten Beispiele angeführt. Dies können bestimmte Organisationen sein, die in diesem Bereich arbeiten und ihre Ergebnisse publizieren oder konkrete Hinweise auf nutzbare Werkzeuge z.B. Emulatoren oder Metadaten Schemata usw. sein. Das entworfene Archiv stellt eine erste Annäherung an ein mögliches Archiv dar. Das OAIS Funktionsmodel wurde wegen seines hohen Bekanntheitsgrades und seines Status als ISO-Standard gewählt. Es sind sicher auch andere Funktionsmodelle möglich.

Siehe Abbildung 15.6.1: OAIS Funktionsmodell

Das Archiv nutzt die Emulation als digitale Erhaltungsmaßnahme. Es wird angenommen, dass das Archiv alle rechtlichen Fragen geklärt hat und die Benutzung der Computerspiele durch die Archivbesucher legal ist.

Ingest/Produzent/Erfassung:

Bevor ein Spiel in das Magazin des Archivs eingestellt werden kann, muss es

von seinem originalen Datenträger auf einen für das Archiv nutzbaren Datenträger überspielt werden. Dieser Vorgang ist mit einem hohen Aufwand verbunden, da die obsoleten Systeme nicht ohne weiteres mit den aktuellen Systemen über kompatible Schnittstellen verbunden werden können. Insbesondere das Auslesen von ROM-Chips erfordert ein hohes Maß an technischer Kenntnis. Teilweise muss auch erst ein Kopierschutz umgangen werden. Da sich fast alle obsoleten Systeme technisch unterscheiden, ist für jede Plattform ein anderes Expertenwissen gefragt. Glücklicherweise wurden diese Arbeiten schon zu weiten Teilen erbracht. Teilweise könnten nahezu komplette Sammlungen fast aller damals gebräuchlichen Systeme aus dem Internet bezogen werden. Ein Nachteil dieser Methode wäre allerdings, dass einem über die Herkunft der bereits migrierten Programme vertrauenswürdige Informationen fehlen. Dies kann zu Problemen führen, wenn die Programme beim Umgehen des Kopierschutzes verändert oder beschädigt wurden. Viele Spiele des C64 Homecomputers, die heute über das Internet im Umlauf sind, sind Produkte der damaligen Softwarepiraterie. Ihr Programmcode wurde von den sogenannten „Crackern", den Knackern des Kopierschutzes, abgeändert. Teilweise wurden die Programme dadurch zerstört. Ein Archiv muss deshalb innerhalb seiner Sammelrichtlinien festlegen, ob es veränderte Programme von unbestimmter Herkunft in seinen Bestand aufnehmen möchte oder nicht.
Die Software Preservation Society, eine Gruppe von Technikexperten für die Migration von Disk Images, akzeptiert nur originale, unveränderte Programme, die mitsamt ihrem Kopierschutz auf neue Datenträger überspielt wurden. Dazu wurde das Interchangeable Preservation Format entwickelt, mit dem sich die Disk Images mit der Hilfe eines Emulators auf einer aktuellen Plattform nutzen lassen. Die Sammlung der SPS umfasst weite Teile der Amiga Spiele.
Eine weitere Frage des Ingests ist: Welche weiteren Informationen werden neben dem Programm noch benötigt, um es später zu archivieren und zu nutzen? Diese Informationen sollten ein Bestandteil des Submission Information Packages (SIP nach der OAIS-Terminologie) sein.
Manche Computerspiele, wie z.B. „Pong", erklären sich von selber. In der Regel benötigt man aber eine Bedienungsanleitung, um ein Spiel zu verstehen. Teilweise enthalten die Anleitungen auch Passwörter, die ein Spiel erst in Gang setzen. Dies war eine häufige Form des Kopierschutzes. Die Bedienungsanleitung ist somit ein fester Teil des Data Objects, dass vom Archiv bewahrt werden muss. Genauso wichtig sind Informationen darüber, welcher Emulator verwendet werden soll. Es wäre auch denkbar, dass der Emulator ein Bestandteil des SIPs ist, wenn das Archiv noch nicht über ihn verfügt. Zur Vollständigkeit trägt auch eine technische Dokumentation der obsoleten Plattform bei, auf der das

Spiel ursprünglich betrieben wurde. Außerdem werden Informationen über den Kopiervorgang, die Herkunft des Spiels und die rechtlichen Bestimmungen benötigt. Um das Bild abzurunden, sollten digitalisierte Bilder der Verpackung, des obsoleten Datenträgers und der Hardware dem Data Object beigefügt werden. Beispiele für solche Scans findet man auf der Web-Seite von ATARI Age oder lemon64.com. Informationen über Langzeitarchivierungsformate für Bilder und Text finden sich in den betreffenden Kapiteln dieses Handbuchs.
Es wäre günstig für ein Computerspielarchiv, wenn die Zeitspanne zwischen der Veröffentlichung eines Spiels und seiner Aufnahme in das Archiv möglichst kurz wäre. Nur solange das Spiel auf seiner originalen Plattform läuft, kann das authentische Verhalten und Look and Feel des Programms durch das Archiv dokumentiert werden. Diese Dokumentation wird später zur Beurteilung des Emulatorprogramms benötigt. Ohne ausreichende Angaben kann später niemand sagen, wie authentisch die Wiedergabe des Spiels mittels des Emulators ist.
Es ist sehr wahrscheinlich, dass sich der Bestand eines Computerspielarchivs nicht allein auf die Spiele als Archivobjekte beschränken kann. Zur technischen Unterstützung müssen, neben den Emulatorprogrammen auch obsolete Betriebssysteme, Treiberprogramme, Mediaplayer usw. archiviert werden.

Archival Storage/Magazin

Die Haltbarkeit der Datenträger hängt von der Nutzung und den klimatischen Lagerungsbedingungen ab. Hohe Temperaturen und hohe Luftfeuchtigkeit können die Lebensdauer eines Datenträgers, ob optisch oder magnetisch, extrem verkürzen. Die Wahl des Datenträgers hängt auch mit der Art des Archivs, seinen finanziellen und räumlichen Möglichkeiten ,sowie den Erwartungen der Nutzer ab.
Sicher ist, dass die Bestände in regelmäßigen Abständen auf neue Datenträger überspielt werden müssen. Dabei sollten die Bestände auf Datenträger des gleichen oder eines ähnlichen Typs überspielt werden, wenn sich das angegebene Verfallsdatum des alten Trägers nähert, oder die Datenträger besonderen Strapazen ausgesetzt waren. Die Bestände sollten auf einen Datenträger eines neuen Typs überspielt werden, wenn der alte Datenträger technisch zu veralten droht. Es ist unwahrscheinlich, dass ein Langzeitarchiv ohne diese beiden Typen von Migration auskommt. Informationen zu den möglichen digitalen Speichermedien finden sie in den betreffenden Kapiteln dieses Handbuchs.
Genauso wie die Datenträger ständig überprüft und erneuert werden, müssen auch die Emulatorprogramme an die sich wandelnden technischen Bedin-

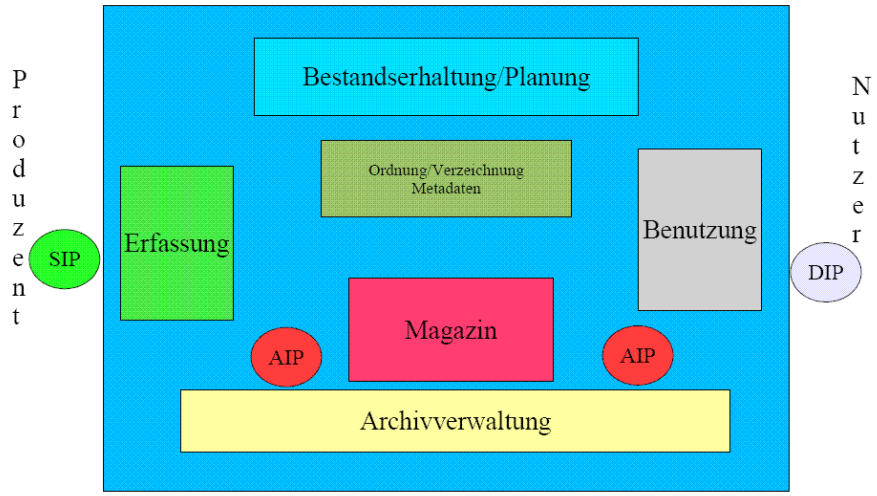


Abbildung 15.6.1: OAIS Funktionsmodell

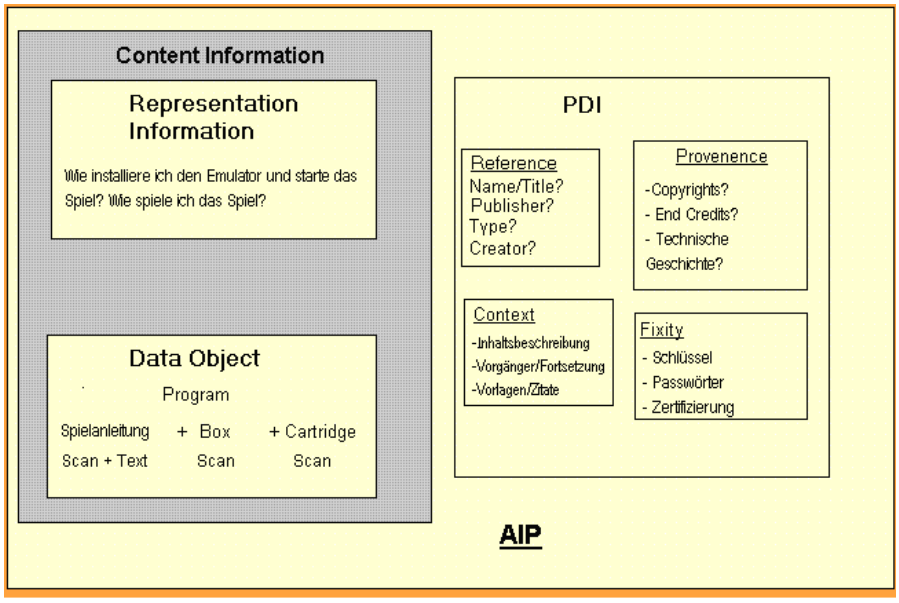


Abbildung 15.6.2: AIP für ein Computerspiel

gungen angepasst werden. Die möglichen Strategien zur Nutzung von Emulatoren entnehmen Sie bitte dem entsprechenden Kapitel des Handbuchs oder den Literaturangaben.
Die Wahl des Emulatorprogramms ist abhängig vom Spiel, dass emuliert werden soll. Ein Spiel, dass von einer Commodore64 Plattform stammt, kann nicht mit einem Emulator verwendet werden, der eine ATARI VCS Plattform emuliert. Zudem sollten Archive bei der Auswahl ihrer Emulatoren weitere Faktoren, wie Benutzungsbedingungen, technische Weiterentwicklung und Hilfestellung durch die Entwicklergemeinde, Leistungsfähigkeit und Authentizität der Darstellung, Einfachheit der Bedienung und Installation, Verbreitung auf verschiedenen Hardware/Software Plattformen usw. bedenken. Es gibt Emulatorprogramme, die von einer internationalen Entwicklergemeinde ständig verbessert und an neue Plattformen angepasst werden. Die weltweit größte Gemeinde hat bisweilen der Multiple Arcade Machine Emulator, der für Arcade-Spiele verwendet wird. Der MAME Emulator unterstützt zurzeit ca. 3.000 Spiele. Ein Ableger von MAME ist das Multiple Emulator Super System (MESS), der Konsolen, Homecomputer, Handhelds und sogar Taschenrechner emuliert. Zur Zeit kann MESS für 442 unterschiedliche Plattformen genutzt werden. Es ist davon auszugehen, dass für nahezu jedes obsolete Spielsystem ein Emulator existiert.

Data Management/Ordnung/Verzeichnis

Es gibt keine moderne Bibliothek ohne Katalog und kein Archiv ohne Findmittel. Auch ein Archiv für Computerspiele braucht ein Verzeichnis. Benötigt werden Metadaten zur inhaltlichen und formalen Erschließung des Bestandes. Bibliotheken nutzen für die formale Erschließung von Computerspielen die Regeln für die alphabetische Katalogisierung für elektronische Ressourcen. Für ein digitales Archiv wäre der Metadatensatz des Dublin Core möglicherweise besser geeignet und unkomplizierter in der Anwendung. Die SPS hat für ihren Katalog einen kleinen Metadatensatz mit den wichtigsten formalen Daten entwickelt.
Die inhaltliche Erschließung erfolgt in der klassischen Bibliothek über Klassifikationen und Systematiken. Einige öffentliche Bibliotheken, die auch Computerspiele in ihrem Bestand führen, haben die verschiedenen Genre, nach denen sich die Computerspiele klassifizieren lassen, in ihre Systematiken eingebaut. Diese Klassifikationen sind aber nicht für ein Spezialarchiv geeignet, das ausschließlich Computerspiele sammelt. Die Klassifikation nach Genres und Subgenres scheint für die inhaltliche Erschließung zumindest der richtige Ansatz zu sein. Es sollte von diesem Punkt aus möglich sein, Spezialsystematiken mit

einer höheren Indexierungsspezifität zu entwickeln, die für ein Computerspielarchiv angemessen sind.
Die inhaltliche und formale Erschließung eines Bestandes findet man auch in der traditionellen Bibliothek. Neu hinzukommen alle Metadaten, die wichtig für den langfristigen Erhalt eines digitalen Objektes sind. Seit neuestem gibt es Metadatenschemata, die diese Informationen erfassen und strukturieren. Bisher werden diese Schemata vor allem für die Langzeitarchivierung von digitalen Texten und Bildern verwendet. Erfahrungen mit der Erfassung von Computerspielen stehen noch aus. Das Metadatenschema PREMIS scheint jedoch ein vielversprechender Kandidat für die Verzeichnung von Langzeitarchivierungsdaten und die Abbildung der Struktur von komplexen digitalen Objekten zu sein.
Ausgehend vom OAIS sollten die Metadaten und das Data Object gemeinsam in ein Archival Information Package (AIP) integriert werden.

Siehe Abbildung 15.6.2: AIP für ein Computerspiel

Alle Informationen des SIP sollen auch im AIP enthalten sein. Als wichtigster Teil des AIP wird die sogenannte Representation Information angesehen. Sie umfasst alle Informationen, die nötig sind, um das Data Object, in unserem Fall das Computerspiel, zu nutzen und zu verstehen. Es wäre demnach ratsam, entweder den entsprechenden Emulator mit Gebrauchsanleitung dort abzulegen oder an dieser Stelle auf den benötigten Emulator zu verweisen. Einige Emulatoren sind schwer zu bedienen. Oft braucht man auch Kenntnisse über die emulierte Plattform, da man sonst nicht weiß, wie das Programm gestartet werden kann. Es ist deshalb ratsam, die nötigen Anweisungen zum Starten des Spiels mittels eines Emulator in einfachsten Schritten, der Representation Information beizufügen.
Die Representation Information ist nicht statisch. Es ist anzunehmen, dass auch die aktuellen Hardware/Software Konfigurationen in absehbarer Zeit veralten. Ebenso wie das Emulatorprogramm muss dann auch die Representation Information an die neuen technischen Bedingungen angepasst werden. Wie bereits erwähnt, scheint PREMIS für diese Aufgabe der beste Kandidat zu sein. Für Archive, die eine größere Freiheit bei der Auswahl ihrer Metadaten benötigen, scheint METS eine gute Alternative zu sein. Beide Metadatenschemata sind in XML-Schemas umgesetzt worden und beanspruchen für sich, OAIS-konform zu sein. Näheres zu PREMIS und METS sowie über Langzeitarchivierungsmetadaten finden Sie in den entsprechenden Kapitel des Handbuchs.

Benutzung

Je besser und genauer die Angaben der Representation Information sind, umso einfacher wird die Benutzung des archivierten Computerspiels. Die Benutzung und die Übermittlung des Spiels hängt hauptsächlich von den Möglichkeiten des Archivs ab. Die Benutzung könnte Online, innerhalb der Räume des Archivs oder durch den Versand eines Datenträgers erfolgen. Neben dem Spiel muss auch der Emulator und die entsprechende Representation Information übermittelt werden. Alle genannten Teile zusammen ergeben das Dissemination Information Package (OAIS-Terminologie). Ein Beispiel für eine benutzerfreundliche Vermittlung wird zurzeit an der Universität Freiburg im Rahmen einer Dissertation entwickelt. Der Nutzer kann einen Emulator und ein Computerspiel über ein Web-Applet in seinem Browserfenster laden und starten. Das Spiel läuft ausschließlich auf seinem Bildschirm, es wird nicht auf die Festplatte des Archivnutzers heruntergeladen.

Zusammenfassung

Eine nachhaltige Archivierung von Computerspielen in einem größeren, öffentlichen, institutionellen Rahmen steht noch aus. Kleinere Organisationen mit dem nötigen technischen Know-how stehen bereit. Technische Arbeitsmittel wie Emulatoren oder Metadatenschemata im XML-Format sind bereits verfügbar. Eine Langzeitarchivierung von Computerspielen ist technisch möglich. Benötigt werden die entsprechenden Mittel, geeignete rechtliche Vorgaben und ein noch zu etablierender Wissenstransfer zwischen den klassischen Institutionen (Bibliotheken, Medienarchive, Museen) und den engagierten kleineren Organisationen mit den technischen Spezialkenntnissen.
Die bisherige Arbeit des Computerspielemuseums Berlin und des Digital Game Archives zeigt, dass ein vielfältiger Bedarf (kulturell, wissenschaftlich) auf der Nutzerseite existiert.

Literatur

Lowood, Henry: Video Games in Computer Space: The complex history of Pong – 2005; in: *Videoludica Vintage* (1971- 1984), eds. Ian Bogost & Matteo Bittanti (Edizioni Unicopli, exp. mid-2007)

Fritz, Jürgen : Was sind Computerspiele? In: Handbuch Medien: Computerspiele: Theorie, Forschung, Praxis/ hrsg. Jürgen Fritz und Wolfgang Fehr – Bonn: Bundeszentrale für politische Bildung Koordinierungsstelle Medienpädagogik; 1997. (S. 81-86)

Huth, Karsten; Lange, Andreas: Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive; In: ICHIM Berlin 04 – Proceedings: 2004; Im Internet: http://www.archimuse.com/publishing/ichim04/2758_HuthLange.pdf (letzter Zugriff 15.10.2007)

Huth, Karsten: Probleme und Lösungsansätze zur Archivierung von Computerprogrammen - Am Beispiel der Software des ATARI VCS 2600 und des C64 – Berlin: Humboldt Universität; 2004: Im Internet: http://www.digitalgamearchive.org/data/news/Softw_Preserv_huth.pdf letzter Zugriff 15.10.2007)

Dooijes, Edo Hans: Old computers, now and in the future – 2000: Im Internet: http://www.science.uva.nl/museum/pdfs/oldcomputers_dec2000.pdf letzter Zugriff 15.10.2007

Computerspiele Museum Berlin: Im Internet: www.computerspielemuseum.de (letzter Zugriff 15.10.2007)

The Digital Game Archive (DiGA): http://www.digitalgamearchive.org/home.php (letzter Zugriff: 15.10.2007)

Lowood, Henry: Playing History with Games : Steps Towards Historical Archives of Computer Gaming - American Institute for Conservation of Historic and Artistic Works. Electronic Media Group: 2004 Im Internet: http://aic.stanford.edu/sg/emg/library/pdf/lowood/Lowood-EMG2004.pdf (letzter Zugriff 15.10.2007)

Internet Archive: Software Archive: Im Internet: http://www.archive.org/details/software (letzter Zugriff: 22.3.2006)

Rulemaking on Exemptions from Prohibition on Circumvention of Technological Measures that Control Access to Copyrighted Works: Im Internet: http://www.copyright.gov/1201/2003/index.html (letzter Zugriff 22.3.2006)

Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1: Blue Book – Consultative Committee for Space Data Systems; 2002: Im Internet http://public.ccsds.org/publications/archive/650x0b1.pdf (letzter Zugriff:15.10.2007)

Software Preservation Society (SPS): Im Internet: http://www.softpres.org/ (letzter Zugriff 15.10.2007)

AtariAge: Im Internet: http://www.atariage.com/ (letzter Zugriff 15.10.2007)

Lemon64: Im Internet: http://www.lemon64.com/ (letzter Zugriff 15.10.2007)

Multiple Arcade Machine Emulator: Im Internet: http://mamedev.org/ (letzter Zugriff: 15.10.2007)

Multiple Emulator Super System: Im Internet: http://www.mess.org/ (letzter Zugriff: 22.3.2006)

Data Dictionary for Preservation Metadata. Final Report of the PREMIS Working Group - Dublin, Ohio: 2005: Im Internet: http://www.oclc.org/research/projects/pmwg/premis-final.pdf (letzter Zugriff: 16.2.2006)

Metadata Encoding and Transmission Standard: Official Website: Im Internet: http://www.loc.gov/standards/mets/ (letzter Zugriff: 22.3.2006)

Rothenberg, Jeff: Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources – Washington D.C.: Council on Library and Information Resources, 1998: S. 18: Im Internet: http://www.clir.org/pubs/reports/rothenberg/criteria.html (letzter Zugriff 22.3.2006)

Welte, Randolph; Suchodoletz, Dirk von: Projekt „Funktionale Langzeitarchivierung: Implementation eines Beispielarchives“ – Universität Freiburg Rechenzentrum, Lehrstuhl für Kommunikationssysteme: Im Internet: http://www.ks.uni-freiburg.de/fla/ (letzter Zugriff 22.3.2006)

16 Lernen und weitergeben – Aus- und Weiterbildungsangebote zur Langzeitarchivierung[1]

Prof. Dr. Achim Oßwald, Prof. Regine Scheffel

1 Zielsetzung

Die im Rahmen des nestor-Projektes entwickelte Website *www.langzeitarchivierung.de* ist eine wahre Fundgrube für Praktiker wie auch für Lehrende. Erstmalig ist damit der Sach- und Forschungsstand zum Thema Langzeitarchivierung in deutscher Sprache breit dokumentiert und zugänglich. Dies hat nicht nur Auswirkungen auf die Wahrnehmung des Themas in Fachkreisen, die dank der parallelen Pressearbeit von nestor erkennbar gestiegen ist. Auch die Sensibilisierung der Öffentlichkeit für Fragen der Langzeitarchivierung ist durch Meldungen, Radio- und TV-Beiträge und die Darstellung von Problemfällen der Langzeitarchivierung wesentlich verbessert worden.
Die nestor-Projektbeteiligten haben zudem mit einer Reihe von Seminarver-

1 Schriftliche Fassung eines von den beiden Autoren gehaltenen Vortrags anlässlich der nestor-Abschlußkonferenz „Den Fortschritt bewahren....". am 19. Juni 2006 in der damaligen Deutschen Bibliothek in Frankfurt am Main. Vgl. a. *http://www.langzeitarchivierung.de/downloads/nestor_2006_06_19_osswald_scheffel.pdf#search=%22Scheffel%20Leipzig%22*

anstaltungen die Grundproblematik sowie den aktuellen Stand der Problemlösungsangebote zielgruppenspezifisch thematisiert und dokumentiert.[2] Die Materialien der Seminarveranstaltungen sind auf den nestor-Seiten weiterhin zugänglich. Zudem wurden die ersten beiden dieser Veranstaltungen in Göttingen mittels Video aufgezeichnet und stehen auf einer 6 Stunden und 35 Minuten umfassenden DVD-ROM zur Verfügung.[3] Dadurch sind auch andere als die klassischen Folien-Unterlagen von Vorträgen abrufbar und insgesamt deutlich bessere Voraussetzung für eine Multiplikation der Erkenntnisse gegeben, als dies sonst bei Forschungsprojekten der Fall ist.

Von all dem profitieren die Aktivitäten der Aus-, Fort- und Weiterbildung immens: Langzeitarchivierung, nestor und die Ergebnisse des Projektes sind dort Themen geworden. Zwar wurde die primäre Fortbildungsnachfrage durch die o. g. nestor-Seminare erst einmal erfüllt, es ist allerdings zu erwarten, dass sie mit steigendem Bewußtsein über das grundsätzliche Problem deutlich zunehmen wird. Hierfür gilt es mittelfristig neue Angebote zu konzipieren, denn nahe liegender weise kann von den bisherigen nestor-Aktiven nicht erwartet werden, daß sie sich diesbezüglich über die Projektlaufzeit hinaus in der Pflicht fühlen.

Im Bereich der Ausbildung von Informationsspezialisten sind in den diversen Fachhochschulen und Universitäten Fragen der Langzeitarchivierung von digitalen Medien in den vergangenen Jahren von verschiedenen Professorinnen und Professoren sowie Lehrbeauftragten als Thema aufgegriffen und in unterschiedlichen Lehrveranstaltungsformen thematisiert worden.

Ziel der nachfolgenden Untersuchung und Dokumentation ist es, eine summarische Bestandsaufnahme hinsichtlich der bisherigen Verortung des Themas Langzeitarchivierung digitaler Medien in den Curricula der einschlägigen Hochschulen[4] vorzunehmen, um auf dieser Grundlage Überlegungen und Konzepte für zukünftige Angebote in der Aus-, Fort und Weiterbildung im Themenbereich Langzeitarchivierung zu entwickeln.

2 Die Zielgruppen des Qualifizierungsbedarfs

Qualifizierungsmaßnahmen müssen inhaltlich hinsichtlich ihrer Zielgruppen und - damit verbunden - hinsichtlich der medialen und didaktischen Aufbereitung ihrer Themen differenziert werden. Als Zielgruppen, für die Qualifizie-

2 Siehe www.langzeitarchivierung.de , -> Veranstaltungen, -> nestor-Seminare: 1. Seminar „Einführung ...“ am 29.11.05 an der SUB Göttingen; 2. Seminar „Archivbereich ...“ am 13.01.06 ebenfalls an der SUB Göttingen; 3. Seminar „Museen ...“ am 13.06.06 in Nürnberg.

3 nestor-Seminare; Göttingen 2006, ISBN 3-938616-41-5.

4 Angebote außerhalb des Hochschulbereichs wurden nicht analysiert.

rungsbedarf besteht, können im Bereich der Langzeitarchivierung mindestens die folgenden Gruppen identifiziert werden:

- Leitungs- und Führungspersonen aus Einrichtungen, für die Langzeitarchivierung relevantes Thema ist oder in absehbarer Zeit werden wird
- Mitarbeiterinnen und Mitarbeiter mit operativen Aufgaben im Bereich Langzeitarchivierung in Einrichtungen, die Langzeitarchivierung durchführen oder an ihr beteiligt sind
- Qualifizierte im Kulturerbe-Bereich (Archiv / Bibliothek / Informationswirtschaft / Museum) und der Informationswirtschaft, die über den state-of-the-art informiert bzw. hierfür sensibilisiert werden sollen und wollen
- Bachelor-Studierende aus dem Kulturerbe-Bereich
- Master-Studierende, die sich für Tätigkeiten im Bereich der Langzeitarchivierung qualifizieren wollen.

Für die Ausrichtung von Qualifizierungsmaßnahmen bedarf es nicht nur entsprechender didaktischer Ausarbeitungen, sondern auch einer vorherigen differenzierten Analyse der jeweiligen Interessen und Bedarfe der jeweiligen Zielgruppen. Dies könnte in einem zukünftigen Projekt im Umfeld von nestor geleistet werden. In den bisherigen Fort- und Weiterbildungsaktivitäten von nestor war dies noch nicht möglich. Hier erfolgte die Aufbereitung der Inhalte bislang lediglich spartenspezifisch.

3 Distributionswege des Wissens über Langzeitarchivierung

Die bisherigen Distributionswege des Wissens in deutscher Sprache zum Bereich Langzeitarchivierung digitaler Medien sind in starkem Maße durch die Bereitstellung von Materialien über die Website des nestor-Projektes geprägt. Dies bezieht sich auf online-Publikationen, aber auch auf die schon erwähnten Video-Mitschnitte von relevanten Fortbildungsveranstaltungen zum Thema. Zu den über *www.langzeitarchivierung.de* angebotenen Materialien sind durch die Kooperation von nestor und dem von der australischen Nationalbibliothek angebotenen Portal PADI (Preserving Access to Digital Information; *http://www.nla.gov.au/padi/*) zu Fragen der Langzeitarchivierung eine Fülle weiterer Materialien für jene in den Blickpunkt gebracht worden, die bislang PADI noch nicht die entsprechende Aufmerksamkeit geschenkt hatten. Verbunden mit dieser und den weiteren, damit angelegten Vernetzungen von Materialien, die über andere Websites weltweit zum Thema bereitgestellt werden, ist in rascher Zeit eine relativ günstige Quellenlage entstanden.

Schon in den ersten Projekten zum Thema Langzeitarchivierung wurden erste e-Learning-Animationen und -Module zu Teilaspekten des Themenbereichs

entwickelt, beispielsweise das Modul „dSEP“ im Rahmen des nedlib-Projektes. Dieses Angebot wird auf dem Archivserver der Deutschen Nationalbibliothek gehostet (vgl. *http://deposit.d-nb.de/netzpub/web_langzeiterhaltung_ep.htm*). Weitere Materialien werden in der Regel auf Projektwebsites bereitgestellt und sind auch durch Suchmaschinen auffindbar. Explizite e-Learning-Angebote zum Thema wie z.B. das der Universitätsbibliothek Cornell[5], sowie das entsprechende Modul im Rahmen des von der UNESCO bereitgestellten IMARK-Modul[6] kommen hinzu.

Und schließlich werden auch auf den jeweiligen Lehrmaterialseiten einzelner Hochschulen bzw. der Lehrenden an Hochschulen Foliensammlungen und vertiefende Lehrmaterialien bereitgestellt.[7] Eine inhaltliche Gesamtschau dieser Angebote aus deutschsprachiger oder auch internationaler Perspektive fehlt bislang.

Die folgende Untersuchung widmet sich der Frage, inwieweit im deutschsprachigen Bereich konkrete Module für Lehrveranstaltungen in den einschlägigen Bachelor- und Master-Studiengängen aus dem engeren informationswissenschaftlichen Bereich sowie dem weiteren Kulturerbe-Bereich den Komplex Langzeitarchivierung aufgreifen, sei es als Thema eines Ausbildungsschwerpunktes, in speziellen Lehrveranstaltungen oder als Thema einzelner Unterrichtseinheiten. Die Untersuchung konzentriert sich also auf eine erste Bestandsaufnahme zu den konkreten Vermittlungsaktivitäten im Rahmen von studien- sowie weiterbildungsrelevanten Veranstaltungen.

4 Bestandsaufnahme: Studienangebote und Studienmodule zum Thema Langzeitarchivierung in Deutschland

Zum genannten Zweck wurde im Juni 2006 eine Analyse von bereits etablierten oder in der Planung befindlichen Bachelor (BA)- und Master (MA)-Studienangeboten[8] an 16 Hochschulen in Deutschland, der Schweiz und Österreich

5 Digital Preservation Management: Implementing Short-term Strategies for Long-term problems; Cornell University Library, 2003; http://www.library.cornell.edu/iris/tutorial/dpm/index.html

6 Information Management Resource Kit (IMARK), http://www.imarkgroup.org/. “IMARK is an e-learning initiative in agricultural information management developed by FAO and partner organizations” (ebd.)

7 Vgl. z.B. Margarete Payer, HdM Stuttgart: http://www.payer.de/digitalebibliothek/digbib02.htm; Achim Oßwald; FH Köln: http://www.fbi.fh-koeln.de/institut/personen/osswald/Material_Osswald/ws05/LZA_digitalerPublikationen_021006_2auf1_sw.pdf; Regine Scheffel, Unterlagen im E-Learning-Portal der HTWK Leipzig.

8 Wir halten es für nur sehr begrenzt sinnvoll, die nunmehr auslaufenden Diplom- und Magis-

durchgeführt. Es handelte sich um die Studienstandorte Berlin, Darmstadt, Düsseldorf, Hamburg, Hannover, Hildesheim, Köln (2x), Konstanz, Leipzig, Potsdam, Regensburg, Stuttgart (2x), Chur (CH), Eisenstadt (AU) sowie Krems (AU).

Analysiert wurden die Studienangebote zunächst auf Grund der Informationen über Studiengänge und Studienpläne auf den Hochschulwebsites. Ergänzend wurden im Bedarfsfall Befragungen per E-Mail durchgeführt, die weitere Einzelheiten zutage förderten und eine Differenzierung bzw. Klarstellung der webbasierten Aussagen ermöglichen.

Die Bestandsaufnahme der Studienangebote ergab, daß Langzeitarchivierung bisher nur einmal als explizites Thema eines Master-Studienangebotes mit dem Fokus Konservierung / Langzeitarchivierung realisiert ist. Es handelt sich dabei um den anwendungsorientierten Master-Studiengang „Konservierung Neuer Medien und Digitaler Information" (Master of Arts) der Staatlichen Akademie der Bildenden Künste Stuttgart.[9] Als postgraduales Studium setzt es einen Hochschulabschluss in Archiv- oder Bibliothekswesen, Informatik, Kunstgeschichte, Medienwissenschaften, Museologie, Restaurierung o. ä. voraus. Inhalte sind Kenntnisse und Fähigkeiten zum langfristigen Erhalt von Kunst, Kultur-, Archiv- und Bibliotheksgut in den Bereichen Fotografie, Video und digitale Information. Unterrichtssprachen sind Englisch und Deutsch. Der erste Kurs ist auf acht Studierende ausgelegt, das kostenpflichtige Studienangebot (1500 € Studiengebühren / Semester) soll später auf 12 Studierende ausgeweitet werden. Anbieter sind die Akademie der Künste in Kooperation mit dem Zentrum für Kunst und Medientechnologie (ZKM Karlsruhe) und weiteren Partnern im In- und Ausland. Abgesichert ist das Angebot durch eine Finanzierungszusage für 5 Jahre seitens des Landes Baden-Württemberg.

Es gibt weiterhin Überlegungen das Themengebiet der digitalen Langzeitarchivierung als kooperativen Vertiefungsschwerpunkt in einem bibliotheks- bzw. informationswissenschaftlichen Master-Studiengang an den Fachhochschulen in Köln und Leipzig anzubieten.

Im gerade angelaufenen europäischen MA-Studiengang „European Multimedia, Arts and Cultural Heritage Studies" an den Universitäten Köln, Coimbra, Lecce und Turku wird das Thema Langzeitarchivierung perspektivisch ebenfalls eine Vertiefungsoption sein.[10]

---

ter-Studiengänge als Bezug zu nehmen, da diese nur noch eine geringe zeitliche Perspektive haben. Für diese bisherigen Studiengänge stellt sich die Situation z.T. schlechter dar.

9 Vgl. für weitere Details http://www.mediaconservation.abk-stuttgart.de/

10 Detailinformationen liegen bislang nicht vor; die Website zu diesem Studienangebot ist in Vorbereitung.

Ansonsten sind Langzeitarchivierung bzw. einzelne Aspekte rund um das Thema und damit auch die nestor-Ergebnisse bislang an einigen Standorten (un)regelmäßig Thema von speziell hieraus ausgerichteten Lehrveranstaltungen der Curricula. Der zeitliche Umfang, in dem Fragen der Langzeitarchivierung thematisiert werden, ist sehr unterschiedlich. Ein kohärentes Curriculum zum Thema ist außer im oben genannten Stuttgarter Studienangebot nicht erkennbar.
Insgesamt ist allerdings ein Fokus auf die Sensibilisierung bezüglich des Themas festzustellen: an 9 Standorten werden diese Themen im zeitlichen Umfang von 2-10 Unterrichtsstunden thematisiert, eingebettet in andere Themen wie z.B. Informationsmanagement / Records Management, Digitales Publizieren / Electronic Publishing, Archivwissenschaft / -typologie oder Museumsdokumentation.
Als Themen mit Bezug zur Langzeitarchivierung nennen die Studiengänge solche, die mit den nestor-Forschungsfeldern korrespondieren. Nach der Häufigkeit ihrer Nennung sind dies:

- Langzeitarchivierung digitaler Daten
- Metadaten
- Archivserver / Open Archival Information System
- Projekte / Infrastruktur für die Langzeitarchivierung
- Persistent Identifier
- Formate
- Open Access
- Rechtliche Aspekte
- Datensicherung.

5 Schlussfolgerungen und Empfehlungen hinsichtlich der Qualifikationsanforderungen an Berufspraktiker

Die bisherigen Vermittlungsaktivitäten sind als ein erster Einstieg zur Sensibilisierung von künftigen Fachleuten aus dem informationswissenschaftlichen oder Kulturerbe-Bereich von Bedeutung. Perspektivisch sollte jedoch eine modular aufgebaute Qualifizierungsstrategie entwickelt werden, die sich an die folgenden drei prioritären Zielgruppen in den Berufsfeldern richtet:

- Entscheidungsträger (E)
- Allgemein Qualifizierte aus dem Kulturerbe-Bereich (Q)
- Mitarbeiterinnen und Mitarbeiter mit Langzeitarchivierungsaufgaben(M)

Die nachfolgende Tabelle konkretisiert, welche zu vermittelnde Inhalte für diese drei Zielgruppen aus Sicht der Autoren sinnvoller weise angeboten werden

sollten. Je nach Interpretation der beschriebenen Inhalte sind hier vermutlich Modifikationen sinnvoll.

Handlungsorientierte Vermittlungsinhalte	**E**	**Q**	**M**
Sensibilisierung + grundlegende Kenntnisse der LZA	X	X	X
Vertiefte Kenntnisse theoretischer Konzepte der LZA (Strategien, Infrastruktur, Sammelrichtlinien, Policies)	X	X	X
Konzeption und Realisierung von Datensicherungs-, Datenrettungs- und Langzeitsicherungsstrategien		X	X
Vertiefte Kenntnisse der Realisierung von Datensicherungs, rettungs- und Langzeitsicherungsstrategien; Archivserverlösungen und deren Durchführung	X		X
Vertiefte Kenntnisse und Anwendungsfertigkeiten bezüglich der Standards, die bei der LZA zur Anwendung kommen			X
Kenntnisse, Fähigkeiten und Fertigkeiten des Daten- und Informations- bzw. Recordsmanagements		X	X
Vertiefte Kenntnis der Informatiklösungen für LZA und deren Anwendung			X
Kenntnis der rechtlichen Aspekte	X	X	X
Vertiefte Kenntnis der rechtlichen Aspekte und ihrer Anwendung			X
Kenntnis der Kostenaspekte	X	X	X

Diese Zusammenstellung deckt sich auch mit der Erwartungshaltung aus der Branche jener Firmen, die sich mit Datenrettung nach Havariefällen befassen.[11]

Auf der nestor-Konferenz vom 19.6.2006 fand diese Aufstellung weitgehende Zustimmung. Es wurden geringfügige Veränderungen vorgeschlagen, die hier eingearbeitet wurden. Allerdings sollte diese Einschätzung durch eine systematische Befragung von Vertretern der Zielgruppen verifiziert werden.

Aus der Aufstellung wird ersichtlich, dass Forschungsfelder in entscheidungs- und handlungsorientierte Vermittlungsinhalte einfließen. Diese Inhalte müssen jedoch kohärent gruppiert und zu Themenmodulen zusammengeführt werden.

Daraus leiten wir folgende **Empfehlungen** ab:

- Kooperative bzw. kollaborative Entwicklungen von Lehreinheiten / Modulen zu den nestor-Forschungsfeldern in didaktisch und medial für ver-

11 Beispielhaft für diese Branche wurde die Firma Ontrack (http://www.ontrack.de/) befragt.

schiedene Zielgruppen aufbereiteter Form
- Konzeption der Module so, dass sie für die Aus- und Weiterbildung sowie in der Fortbildung genutzt werden können, z.B. durch Einbettung in bzw. Umsetzung als e-Learning-Applikationen
- Vermittlung von best-practice-Lösungen der verschiedenen Konzepte auf internationaler, nationaler, regionaler und lokaler Ebene (inkl. der Verknüpfung mit entsprechenden Anschauungsanwendungen)
- Vermittlung der gängigen nationalen und internationalen Normen / Standards an praktischen Beispielen.

Die Angebote sollten - und dies wäre eine deutliche Veränderung zu den bisherigen Aus- und Weiterbildungsangeboten - nicht vorwiegend theorielastig sein, sondern auch praktische Übungen einschließen, z.B.:
- Praktische Übungen zum Handling von Daten in unterschiedlichen Formaten
- Praktische Übungen mit Datensicherungssystemen
- Praktische Übungen in OAIS-basierten Testumgebungen.

Die genannten Überlegungen und Konsequenzen aus der hier vorgelegten summarischen Bestandsaufnahme lassen deutlich werden, daß eine Weiterführung der nestor-Aktivitäten notwendig ist. Auf diese Weise würden systematische Aktivitäten möglich gemacht, um koordinierte Qualifizierungsstrukturen bei bzw. mit Partnern (Hochschulen / Fortbildungseinrichtungen des Kulturerbe-Bereichs) aufzubauen und somit den Erfahrungs- und Erkenntnistransfer aus den bisherigen nestor-Projektaktivitäten sicher zu stellen!

6 Relevante Internetadressen

Der Zugriff auf die nachfolgend genannten Webseiten erfolgte zuletzt in der ersten Oktoberhälfte 2006.

6.1 Studienangebote der folgenden Hochschulabteilungen wurden ausgewertet:

In alphabetischer Reihenfolge der Städtenamen:

Berlin

Humboldt Universität, Institut für Bibliotheks- und Informationswissenschaft
http://www.fbiw.hu-berlin.de/startseite/willkommen/

Chur

Hochschule für Technik und Wirtschaft Chur, Arbeitsbereich Informationswissenschaft

http://www.informationswissenschaft.ch

Darmstadt

Hochschule Darmstadt, Fachbereich I nformations- und Wissensmanagement
http://www.iuw.h-da.de/

Düsseldorf

Heinrich-Heine-Universität Düsseldorf, Philosophische Fakultät**,** Abteilung für Informationswissenschaft am Institut für Sprache und Information
http://www.phil-fak.uni-duesseldorf.de/infowiss/content/studiengaenge/index.php

Hamburg

Hochschule für angewandte Wissenschaften Hamburg, Fakultät Design, Medien und Information - Department Information
http://allekto.bui.haw-hamburg.de/studieren/studienmaterialien.php

Hannover

Fachhochschule Hannover, Fachbereich Informations- und Kommunikationswesen
http://www.ik.fh-hannover.de/de/studium/

Hildesheim

Stiftung Universität Hildesheim, Fachbereich III, Institut für Angewandte Sprachwissenschaft
http://www.uni-hildesheim.de/de/studiumifas.htm

Köln

Fachhochschule Köln, Fakultät für Informations- und Kommunikationswissenschaften, Institut für Informationswissenschaft
http://www.fbi.fh-koeln.de/studium/studium.htm

Köln

Universität zu Köln, Schwerpunkt Medienkulturwissenschaft
http://www.medienkulturwissenschaft.uni-koeln.de/zfmk.html

Konstanz

Universität Konstanz, Informatik & Informationswissenschaft
http://www.inf.uni-konstanz.de/Lehre/IE/ie.html)&

Krems

Donau-Universität Krems, Department für Wissens- und Kommunikationsmanagement
http://www.donau-uni.ac.at/wuk/bim

Leipzig

Hochschule für Technik, Wirtschaft und Kultur, Fachbereich Medien
http://www.fbm.htwk-leipzig.de/

Potsdam

Fachhochschule Potsdam, Fachbereich Informationswissenschaften
http://informationswissenschaften.fh-potsdam.de

Regensburg

Universität Regensburg, Informationswissenschaft
http://www-iw.uni-regensburg.de/mamboiw/index.php?option=com_content&task=view&id=40&Itemid=69

Stuttgart

Staatliche Akademie der Bildenden Künste Stuttgart
http://www.mediaconservation.abk-stuttgart.de/

Stuttgart

Hochschule der Medien, Fakultät Information und Kommunikation
http://www.hdm-stuttgart.de/studienangebot/information_und_kommunikation

6.2 Weitere einschlägige Quellen und Adressen

- Digital Preservation Management: Implementing Short-term Strategies for Long-term problems; Cornell University Library, 2003; http://www.library.cornell.edu/iris/tutorial/dpm/index.html
- Information Management Resource Kit (IMARK), http://www.imarkgroup.org/
- www.langzeitarchivierung.de , -> Veranstaltungen, -> nestor-Seminare
- Oßwald, Achim; Scheffel, Regine: Lernen und weitergeben - Aus- und Fortbildungsangebote zur Langzeitarchivierung;
  Folien des Vortrags der beiden Autoren anlässlich der nestor-Abschlusskonferenz „Den Fortschritt bewahren...“. am 19. Juni 2006 in der dama-

ligen Deutschen Bibliothek in Frankfurt am Main.
http://www.langzeitarchivierung.de/downloads/nestor_2006_06_19_osswald_scheffel.pdf#search=%22Scheffel%20Leipzig%22